THE NEW YORK ACADEMY OF SCIENCES
(LYCEUM OF NATURAL HISTORY, 1817–1876)

OFFICERS, 1945

President: WALTER H. BUCHER

Vice-Presidents: MARSHALL KAY; RAYMUND L. ZWEMER; ANNE ROE; HORTENSE POWDERMAKER; JOSEPH S. FRUTON; RAYMOND B. MONTGOMERY

Recording Secretary: MICHAEL HEIDELBERGER

Corresponding Secretary: H. HERBERT JOHNSON

Treasurer: DONALD BELCHER

Librarian: BARNUM BROWN

Editor: ROY WALDO MINER

Elected Councilors:

1943–1945: CHARLES M. BREDER, JR.; HERBERT J. SPINDEN; HAROLD E. VOKES

1944–1946: VICTOR K. LAMER; THEODORE SHEDLOVSKY

1945–1947: RALPH H. CHENEY; ROBERT CUSHMAN MURPHY; HERBERT F. SCHWARZ

Finance Committee: HARDEN F. TAYLOR (*Chairman*); HARRY B. VAN DYKE; ADDISON WEBB

OFFICERS OF SECTIONS

GEOLOGY AND MINERALOGY

Chairman: MARSHALL KAY. *Secretary:* RALPH J. HOLMES

BIOLOGY

Chairman: RAYMUND L. ZWEMER. *Secretary:* ROSS F. NIGRELLI

PSYCHOLOGY

Chairman: ANNE ROE. *Secretary:* EMILY T. BURR

ANTHROPOLOGY

Chairman: HORTENSE POWDERMAKER. *Secretary:* GORDON F. EKHOLM

PHYSICS AND CHEMISTRY

Chairman: JOSEPH S. FRUTON. *Secretary:* RAYMOND M. FUOSS

OCEANOGRAPHY AND METEOROLOGY

Chairman: RAYMOND B. MONTGOMERY. *Secretary:* JOHN C. ARMSTRONG

The Sections of Geology and Mineralogy, Biology, Psychology, and Anthropology meet in regular session on Monday evenings at 8:00 o'clock from October to May, inclusive. The Sections of Physics and Chemistry, and Oceanography and Meteorology hold two-day conferences at irregular periods. All meetings of the Academy are held at The American Museum of Natural History, Central Park West at 79th Street, New York, N. Y.

Annals of The New York Academy of Sciences
Volume XLVI, Art. 1, Pages 1–126
June 15, 1945

ANIMAL COLONY MAINTENANCE*

By

Edmond J. Farris, F. G. Carnochan, C. N. W. Cumming, Sidney Farber, Carl G. Hartman, Frederick B. Hutt, J. K. Loosli, Clarence A. Mills, and Herbert L. Ratcliffe

---

CONTENTS

* This series of papers is the result of a conference on Animal Colony Maintenance held by the Section of Biology of The New York Academy of Science, November 10 and 11, 1944.
Publication made possible through a grant from the income of the Conference Publications Revolving Fund.

INTRODUCTION TO THE CONFERENCE ON ANIMAL COLONY MAINTENANCE

By Edmond J. Farris

The Wistar Institute, Philadelphia, Pa.

It is indeed a pleasure to welcome you here on behalf of The New York Academy of Sciences, and it is my privilege to call this meeting to order. I wish particularly to express my appreciation and gratitude to the men who have so kindly contributed the formal papers for the program. The speakers are distinguished in their respective fields, and their messages will prove stimulating to this conference for frank discussion.

This conference results from the fact that there are many unsettled problems regarding maintenance of laboratory animals. I recall in my undergraduate student days in physiology (and I might add in my days of great inexperience with animals), after spending several hours preparing for and conducting a "type" mammalian experiment, my result hardly ever agreed with the usual textbook picture. This was common classroom experience I inquired of the professor, "What's wrong?" The stock response was given, "If it's not your technique, it's the animal." As a graduate student, I had opportunity to observe some of the methods of maintaining laboratory animals, and my retort courteous to my professor, today, would be "Check the laboratory animal quarters. Check the animal caretaker and his methods. Know the animal's history and background." I am sure all the fine equipment on display in many laboratories will not aid in solving biological problems until satisfactory animals for experimental purposes are available.

As biologists, we are more or less consciously substituting a laboratory animal for man—in effort to solve problems presented by man.

Animal experimentation is usually justified on the grounds that the results may be carried over to man, who is an inconvenient experimental animal.

In the past, this was done in qualitative terms merely. For example, a given diet proved inadequate for the growth of a mammal, and we expected it would prove inadequate for man, too. Today, anatomical, physiological, and nutritional research has become quantitative in character. In any such investigations, a relatively uniform animal is desirable. A geneticist prefers mutations and variations, and rightfully so for his purpose. Yet, it is evident that healthy, clean, vigorous stock is essential for accurate research. A standardized animal is needed, for a standardized animal is to the biologist what the pure chemical is to the chemist.

Certainly, to secure such an animal is hardly possible, but, by proper precautions, a nearly standardized animal or group of animals for research purposes should be the aim of most laboratories.

As biologists, we know that the lower mammals differ from higher in being less able to regulate their physiological processes. They are more directly responsive to changes in environment. Respiration, pulse rate and temperature changes are examples of this lack of regulation. It is familiar to us, also, that appearance of a stranger in the colony disturbs animals, as evidenced at the dairy or chicken farm, where the output of milk and eggs is lowered. The animals, though long domesticated, are readily disturbed.

The animal is a very sensitive bit of apparatus. As the late Dr. Henry Donaldson once described the rat, it has more tricks than a string galvanometer, and must be treated with care and consideration. I recall that, on one occasion, simply moving rats from one building to another prevented breeding for several months.

The animal should be contented and happy, and this condition is not easy to attain, but is necessary.

The papers presented at this conference cover the problems involved in animal colony maintenance, and deal especially with six topics: namely, genetic purity; the mating of mammals; feeding laboratory animals; infectious diseases of laboratory animals; temperature control; and, finally, financing and budgeting.

GENETIC PURITY IN ANIMAL COLONIES

By F. B. Hutt

Department of Poultry Husbandry, Cornell University, Ithaca, N. Y.

INTRODUCTION

This symposium arises from recognition of the fact that, while aims and methods differ, there are some problems common to the maintenance of all experimental animals, whether the colony be a half-dozen cows in the feed-lot or a few hundred white rats in the wire-bottomed cages of the nutritionist. The guinea pig may be to the bacteriologist merely a test-tube, to the student of tick-borne diseases a culture medium, and to the geneticist a source of intriguing mutations; but all three can profitably compare notes on its maintenance, even though they view it through spectacles of different colors.

It seems desirable to state at the outset that this discussion of genetic purity and any suggestions made in this paper are not directed to the speaker's fellow-geneticists This does not imply that there are no problems in their animal colonies or that they are particularly impervious to suggestions It is merely that their conception of an animal colony is so different from that of other biologists

The physiologist, the pathologist, and particularly the nutritionist want a steady supply of healthy animals with a minimum of variability. The anatomist appreciates variability a little more. He is accustomed to determining the range in size of an organ, not merely its mean. One hears frequently of the need for greater uniformity in the animals used for experiments in nutrition. The widespread utilization for this purpose of the inbred rats of the Wistar strain indicates the desire, not only to reduce variability to a minimum within any one experiment, but also to prevent as far as possible undue discrepancies between results at different institutions. When white rats of one strain gain only 14 grams in five weeks on a diet low in thiamine and those of another strain gain over twice as much on the same diet, the interpretation of results is likely to be confused unless the two strains are in one laboratory, and the difference, therefore, properly attributed to genetic differences between strains, as was done by Light and Cracas (1938)[1], rather than to variation in diets or in environment.

The geneticist, on the other hand, thrives on variability. It is his stock-in-trade. The recalcitrant rat that lingers on long after its orthodox litter-mates have terminated their abbreviated careers on the

[1] Light, R. F., & L. J. Cracas. Science 87: 90. 1938

diet lacking vitamin Q is merely a statistical nuisance to the nutritionist, but, to the geneticist, it is the potential progenitor of a race able to manage nicely with much less vitamin Q than millions of rats that are less fortunate in the matter of genes. Sometimes, the geneticist may, by inbreeding, eliminate all innate variability except the two or three mutations currently under study, an accomplishment that is particularly easy if his animal colony consists of a few thousand *Drosophila* in pint milk bottles. In other cases, he may induce mutations by heat, by x-rays, or by other means, or he may accumulate naturally occurring mutations to make his various strains larger or smaller, blacker or whiter, more resistant or more susceptible to this or that disease. All of these things he is able to do without special advice or encouragement on this occasion, but he will be interested in what his fellow biologists can tell him about feeding, housing, management and other topics of this symposium.

GENETIC DIFFERENCES BETWEEN INDIVIDUALS

It is a common belief that genetic variability is eliminated by using litter mates as experimental and control animals. While this is true of litter mates in a highly inbred line, it does not hold good for others. Litter mates may differ as much as any two siblings born several weeks, months or years apart. Because of this, and because mutations occur even in inbred lines, it seems desirable to show by a few examples how much a mutation in a single gene locus of the thousands that determine an animal's inheritance may affect its anatomy, physiology, or chances of survival. Because some biologists still think that genes cause only such "sports" as freaks of hair color, of hair form, of eye-color, or other inconsequential mutations, but have little or no effect on fundamental physiology, these examples are chosen to refute that view.

Consider the effects of the gene A^y, one of a series of multiple alleles affecting hair color in the house mouse. In the almost forty years that have elapsed since Cuenot reported that yellow mice do not breed true, it has been clearly shown that this gene is lethal to the homozygote, which dies *in utero*. The result is a ratio of 2 yellow (heterozygotes) to 1 non-yellow. The heterozygotes are characterized by (1) adiposity (Danforth, 1927),[2] (2) a sub-normal metabolism (Benedict and Lee, 1936),[3] and, as Castle (1941)[4] has shown, (3) a slightly greater body

[2] **Danforth, C. H.** J. Hered. **18:** 153–162. 1927
[3] **Benedict, F. G., & E. C. Lee.** Annales de Physiol. **12:** 983–1064. 1936.
[4] **Castle, W. E.** Genetics **26:** 177–191. 1941.

size, apart from their excessive fat. In addition to all these effects, or because of some of them, the yellow mice are (4) less susceptible to spontaneous mammary carcinoma than are their black or brown litter mates (Little, 1934).[5] It is probable that, with further study, still more peculiarities of physiology will eventually be added to the multiple effects of this gene that causes yellow hair.

In the Frizzle fowl, the feathers curl back toward the head. Studies by Landauer and Dunn (1930)[6] and by Hutt (1930)[7] showed that this is caused by an autosomal dominant gene. When frizzled fowls of show-room type are mated together, they yield progeny that are so extremely frizzled as to appear woolly, others of standard type, and some not frizzled at all, the three types in a ratio of 1:2:1. The extremely frizzled birds eventually become more or less bare as their feathers break off. Landauer (1942)[8] and his co-workers have reported that these homozygous Frizzles differ from normal fowls in (1) viability, (2) rate of growth, (3) age of sexual maturity, (4) metabolism, (5) rate of heart beat and size of heart, (6) frequency of different types of blood cells, (7) size of endocrine glands, and still other ways. Similarly a gene for nakedness (Hutt and Sturkie, 1938)[9] is lethal to about half of the affected chicks before hatching, to half of those that do hatch in the first six weeks, and leaves the remainder with physiological handicaps similar to those of the homozygous Frizzles. There are at least three mutations that cause different degrees of nakedness in the mouse, and similar genes have been studied in rats, rabbits, cats, dogs and other animals.

A good example of a simple recessive mutation that prevents the functioning of an important endocrine gland is the dwarfism discovered by Snell (1929)[10] in the house mouse, and shown by Smith and MacDowell (1931)[11] to be caused by failure of the anterior lobe of the pituitary to secrete the growth-promoting hormone. This mutation is particularly interesting, because, while it inhibits that function of the anterior pituitary, it does not affect its secretion of the gonadotropic hormone. Manifold effects of the mutation, (as reviewed by Grüneberg, 1943)[12] include (1) endocrine malfunction, (2) an adult weight about one-quarter of normal, (3) complete sterility in both sexes, (4)

[5] **Little, C. C** J. Exper. Med. **59**: 229–250. 1934.
[6] **Landauer, W., & L. C. Dunn.** J. Hered. **21**: 290–305. 1930.
[7] **Hutt, F. B.** J. Genetics **23**: 109–127. 1930.
[8] **Landauer, W.** Biol Symposia **6**: 127–166. 1942.
[9] **Hutt, F. B., & P. D. Sturkie.** J. Hered. **29**: 370–379. 1938.
[10] **Snell, G. D.** Proc. Nat. Acad. Sci. **15**: 733–734. 1929.
[11] **Smith, P. E., & E. C. MacDowell.** Anat. Rec. **46**: 249–257. 1931.
[12] **Grüneberg, H.** The Genetics of the Mouse. xii & 412 pp. Cambridge Univ. Press. 1943.

histological abnormalities in the anterior lobe, (5) small and abnormal thyroids, (6) subnormal metabolism, (7) infantile structure of the thymus, adrenals and gonads, and (8) subnormal viability. In the rat, there is a simple recessive dwarfism which, though making mature dwarfs only half the size of their normal litter mates, resembles that in the mouse in causing complete sterility.

All dwarfism is not associated with endocrine malfunction and physiological abnormalities. The speaker is now studying a type of dwarfism in the fowl in which the affected birds apparently lay as many eggs and reproduce as well as their normal sisters, even though almost half the normal size. It is particularly interesting because it is sex-linked and recessive. This means that a male, heterozygous for the gene, will produce daughters half of which are normal and half dwarfed, even when he is mated with unrelated normal hens The mutation was sent in for study by the breeder in whose flock it first appeared. He, like those who participate in this symposium, wished to maintain the genetic purity of his colony.

Particularly interesting and important to the man trying to maintain some uniformity in his animals, are the genes that cause some abnormality of metabolism. A good example is the inability of Dalmatian coach dogs to break uric acid down to allantoin as is done by most other dogs. As a result, they excrete over four times as much uric acid per kilogram of body weight per day as do other dogs. The difference, as Trimble and Keeler (1938)[13] have shown, is determined by a recessive mutation and is not linked with the spotting that characterizes the breed. Another genetic abnormality of protein metabolism causes the accumulation of the pigment, porphyrin, in the bones and teeth, its excretion in the urine and extreme photosensitivity. It is a recessive mutation in cattle (Fourie, 1939)[14] and apparently also in man. One mutation prevents the oxidation of phenylalanine beyond the stage of phenylpyruvic acid, and another causes the incomplete katabolism of tyrosine, with resultant excretion of homogentisic acid and the condition known as alcaptonuria. It seems probable that an abnormality in the metabolism of fats is responsible for the accumulation of sphingomyelin in the tissues of persons afflicted with amaurotic idiocy, a condition which is a simple recessive character. Although these last three examples of mutations affecting metabolism are so far known only in man, there is no reason why they should not be found in any other mammal.

[13] **Trimble, H. C., & C. E. Keeler.** J. Hered. 29: 280–289. 1938.
[14] **Fourie, P. J. J.** Onderstepoort J. Vet. Sci. and Animal Ind. 13: 383–398. 1939.

Resistance to disease is usually dependent upon more than one gene, but there is at least one case on record in which a simple recessive mutation was responsible for the loss of an entire strain of animals. This happened in the guinea pigs that were genetically unable to form blood complement (Rich, 1923).[15] These were demonstrated experimentally to be extremely susceptible to infection with *B. cholerae suis* and also proved so susceptible to natural infections that attempts to preserve them even by distribution in several different laboratories were of no avail.

Finally, no mutations have as disastrous effects as the lethal genes. In the homozygous condition, these genes may terminate life before birth, immediately after it, or at various stages up to maturity, or even in later life. Sometimes, their effects are so great that even the heterozygote is visibly affected, as in the Creeper fowl, Dexter cattle, and yellow mice. More often, the carrier of the gene is apparently fully normal and detectable only by a breeding test. Since these genes lessen the efficiency of reproduction, they are particularly important to all animal breeders. Apparently, they are abundant in all species. Ten years ago, a review of lethal mutations then known in domestic animals (Hutt, 1934)[16] listed only 31, but many more have been discovered in the intervening decade, and a recent check-list compiled by Lerner (1944)[17] includes 70 lethals in farm animals, with no less than 25 in cattle alone. Undoubtedly, more will be discovered. Lethal genes should cause little trouble in the strains of laboratory animals that are already highly inbred, but otherwise their presence is to be generally expected, and particularly so where small litters or high post-natal mortality prevail.

GENETIC DIFFERENCE BETWEEN BREEDS AND STRAINS

The foregoing examples were chosen to show that single gene substitutions (or deletions, which may be responsible for some lethal mutations) may have very great effects. Even more important, so far as physiological characters are concerned, are the multiple factors, each exerting a small effect, but in their cumulative action responsible for big differences in form and function. They are important to the breeder of live stock because they affect rate of growth, body size, the utilization of feed, the capacity to produce milk or eggs and other characters of economic value. They are important to the biologist for these

[15] **Rich, F. A.** Vermont Agric. Exper. Sta. Bull. **230**: 24 pp. 1923.
[16] **Hutt, F. B.** Cornell Vet. **24**: 1–25. 1934.
[17] **Lerner, I. M.** J. Hered. **35**: 219–224. 1944.

same reasons, but also because they differentiate strains with respect to nutritional requirements, endocrine functions, resistance to disease, and other important characteristics, and, by so doing, cause discrepancies between experiments and disagreements between results in different laboratories.

A few examples will illustrate the importance of these differences. Considering nutritional requirements, Light and Cracas (1938)[1] found significant differences in the rates of growth of three strains of rats on a diet deficient in thiamine. After pointing out that variations in the factor for the conversion of Sherman-Chase units to International Units could be explained by differences between strains, they pointed out that each laboratory should determine a conversion factor for its particular strain of rats. Similarly, White Leghorns require less thiamine than Rhode Island Reds (Lamoreux and Hutt, 1939).[18] On a diet containing 30 parts per million of manganese, Gallup and Norris (1939)[19] found White Leghorns to be completely free of abnormalities, while, in New Hampshires, 7 per cent of one strain and 18 per cent of another developed perosis, a condition indicating an unusually high requirement of manganese for bone development. Gowen (1936)[20] found that strains of rats differed greatly in their requirements of vitamin D, and white pigs, according to Johnson and Palmer (1939),[21] can store, during exposure to sunshine, a reserve of vitamin D sufficient to last about twice as long after confinement as the reserve stored by black pigs under the same conditions.

Apart from requirements of specific vitamins or minerals, Morris, Palmer and Kennedy (1933)[22] were able to establish by selection two strains of rats, one of which was 40 per cent less efficient than the other in the utilization of food. Most of the differentiation was accomplished with only six generations of selection.

Students of cancer have provided extensive evidence that strains of mice differ not only in susceptibility to a single type of tumor, whether spontaneous or induced, but also in susceptibility to different kinds of tumors. Strains of animals differ also in susceptibility to bacterial diseases. An important point demonstrated by Gowen (1933)[23] is that genetic resistance to disease is more likely to be specific than general. Thus, Schott's mice, bred for resistance to *S. aertrycke*, could

---

[18] **Lamoreux, W. F., & F. B. Hutt.** J. Agric. Res. 58: 307–316. 1939.
[19] **Gallup, W. D., & L. C. Norris.** Poultry Sci. 18: 76–82. 1939
[20] **Gowen, J. W.** Genetics 21: 1–23. 1936
[21] **Johnson, D. W., & L. S. Palmer.** J. Agric. Res. 58: 929–940. 1939.
[22] **Morris, H. P., L. S. Palmer, & C. Kennedy.** Minn. Agric. Exper. Sta., Tech. Bull. 92, 56 pp. 1933.
[23] **Gowen, J. W.** Quart. Rev. Biol. 8: 338–347. 1933.

easily withstand a dose of that organism that was fatal to all mice of the Sil and W. F. lines, but the latter were much more resistant to the virus of pseudorabies than were the Schott mice. Against the antigenic poison, ricin, the Sil strain was most resistant, the W. F. line most susceptible and the Schott mice intermediate.

THE REDUCTION OF GENETIC VARIABILITY

Having thus emphasized the extent to which a single mutation with great effects or many little mutations with cumulative effects may upset the efficient conduct of any experiment with animals, some reassurance is desirable. The fact that many laboratories are operating animal colonies successfully without too much genetic sabotage provides that assurance. The question is merely, "How can the genetic variability in my stock be reduced and kept at a minimum?"

Breeds, Varieties and Strains. Variability is lowest in a "pure line." To Johanssen a pure line was a group of organisms descended from a single ancestor, all identical in genotype and homozygous with respect to every pair of alleles. A clone, such as a patch of potatoes, all reproduced vegetatively from a single tuber, could also be considered a pure line, even though heterozygous in some genes. Such a degree of genetic purity, though readily attained in Johanssen's self-fertilized beans, is not likely to be established or maintained in many animal colonies. For one thing, while almost complete homozygosity can be attained by eight generations of self-fertilization, it takes over sixteen to accomplish the same result with brother x sister matings, which afford the most intense inbreeding that is possible in animals. For another, the occurrence of new mutations tends to increase heterozygosity in any stock, even one that may once have been highly inbred. It seems probable also that, in laboratories where most of the animals are wanted for other experiments, there may be somewhat more difficulty in maintaining the brother x sister matings which King (1919)[21] used so effectively to reduce variability within inbred families of the justly famous Wistar rats.

It is questionable, however, if such a high degree of genetic purity be necessary in most colonies and, as will be shown later, there are advantages in having some genetic variability in one's experimental animals, at least in some kinds of research. Among the domestic animals, there are at hand large colonies in which different degrees of homozygosity have already been established by the breeders, and from

[21] King, H. D. J. Exper. Zool. 29: 71-111. 1919.

which experimental animals may be drawn with some assurance of consistent performance.

For example, the poultry breeder unconsciously recognizes three different degrees of homozygosity in the population of domestic fowls. Firstly, there are the breeds, usually identified by the names of the places in which they were developed, such as Leghorn, Rhode Island Red, Sussex, Minorca, etc. While these were differentiated primarily on the basis of conformation and color, they also differ in many important physiological traits of which the founders of the breeds were quite unaware (Hutt, 1941).[25] Secondly, there are the varieties within a breed, usually distinguished by different colors, but sometimes by other mutations. Finally, the breeder recognizes within a variety different strains, duly tagged with the name of the breeder who developed them, and believed, usually rightly, to differ in such important characteristics as the ability to lay eggs, to withstand disease, or to win blue ribbons at the Boston poultry show.

Each of these classes,—breed, variety and strain—represents a different degree of inbreeding, and consequently of genetic purity or homozygosity, that narrows down the range of variation. To illustrate,—few Leghorns exceed 5 lbs. in weight at maturity, all lay white-shelled eggs, and all are characterized by yellow shanks. This has resulted merely from the collection of certain genes that determine these characteristics and the exclusion from the breed of genes that make bantams, giants, brown shells or slaty shanks. It is the process of selection, but it is also a measure of inbreeding. Genetic variability is not reduced very much further by the distinguishing characteristic of the White Leghorns, a dominant gene, *I*, which inhibits the formation of melanin. However, once that color variety is set apart, an additional degree of inbreeding is assured, because selection of breeding-stock is limited to the white birds. Finally, the development of strains within a variety is usually carried out by selection within a single flock, and the inbreeding process is thus carried one step further.

Breeds, varieties, strains or races are available in most of the animals that have been long domesticated and in the laboratory animals that reproduce rapidly. They represent different degrees of inbreeding,—of genetic purity—already available for the laboratory worker. Although genetic variability is consistently reduced with each of these steps, much heterozygosity still remains. Thus, all attempts to establish an inbred strain of fowls by brother x sister matings have thus far

[25] Hutt, F. B. Proc. Seventh Internat. Congr., Genet. (Edin. 1939): 156–157. 1941.

failed because of the many lethal genes present in the heterozygous condition in the foundation stock. However, this method has been successfully used to establish cultures of *Drosophila* that are either "pure lines" or as close to it as one may go in animals. Among mammals, brother x sister matings for many generations have yielded the highly homozygous Wistar rats, the guinea pigs of the United States Department of Agriculture and the dilute brown mice developed by Little which have been so widely used for the study of neoplasms.

FURTHER INBREEDING. The extent of inbreeding desirable will vary with the species and the type of experiment. Workers with *Drosophila* and with mice have little difficulty in maintaining strains that are completely homozygous, or nearly so, by continuous brother x sister matings. The same could probably be done with rats, rabbits and guinea pigs, provided that enough animals could be reared to prevent the colony from being closed out by lethal genes, that inevitably show their effects as inbreeding progresses. Such a risk would be minimized by starting with animals already somewhat inbred and there is greater chance of success if special care be taken to select breeding stock from large litters that show a minimum of undesirable defects. After the first generation, several different families must be maintained at least until some of them have weathered the first six generations. This is the most critical period, because it is that in which the greatest reduction of heterozygosity occurs (FIGURE 1). The process of inbreeding

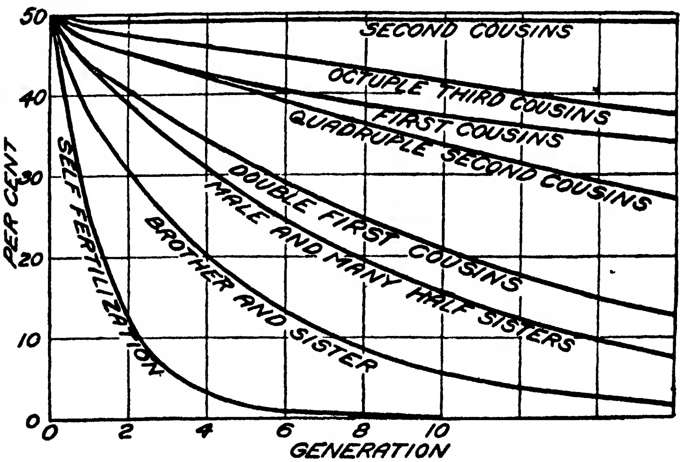


FIGURE 1 Rates at which heterozygosity is reduced with different degrees of inbreeding (From Wright, in U.S.D.A. Tech. Bull. 1121).

will be accompanied by differentiation into families, some of which will have undesirable traits necessitating their elimination, and the breeder must expect a considerable loss of stock before he can establish a strain that is satisfactory.

If it were not for the occurrence of new mutations, inbred lines like these could be kept homozygous in all genes. Even allowing for new mutations, the use of such stocks should greatly reduce variability in most experiments. After 25 generations of brother x sister matings in the Wistar rats, King (1919)[24] found the coefficients of variability in body weight at a given age to be only 10.9 and 8.5 per cent in the inbred males and females, compared with coefficients of 18.0 and 14.8 in controls. A similar reduction in variability with inbreeding is to be expected in almost any character, except those in which variation induced by the environment is much greater than that controlled by genes.

In species or strains in which brother x sister matings have been tried without success, it is probable that some satisfactory degree of homozygosity can be established with a less intense form of inbreeding. Half a dozen attempts to establish an inbred line of fowls by brother x sister matings failed, but at the Iowa Experiment Station a strain of highly inbred Leghorns was produced by using birds more distantly related. This takes a longer time, but lessens the risk of complete failure. Wright (1931)[26] calculated that, whereas the residual heterozygosity is reduced 19.1 per cent with each generation of brother x sister matings, it is reduced only 11 per cent if a single male can be mated in each generation with a large number of his half sisters The rates at which this system of mating and others decrease the heterozygosity are shown in FIGURE 1.

Finally, some reduction of variability can be induced merely by breeding within one colony and not introducing any new blood. Contrary to an opinion commonly held, this does not result in rapid inbreeding, if the breeding stock be chosen at random and if several different sires are used in each generation. With such a system, according to Wright (1931),[26] the heterozygosity is reduced in each generation by approximately $\frac{1}{8N}$, where N is the number of males used. provided that the number of females is "unlimited." Although the use of 30 or 40 females per generation is hardly "unlimited," their number is relatively unimportant to the formula if they are far in excess of the number of males, as is usually the case. Thus, if 10 males be used per

[26] Wright, S. Genetics 16: 97–159. 1931.

generation, the remaining heterozygosity is reduced only about 1.25 per cent, but if only three are used, the figure is about 4.2 per cent. Such a system is not likely to cause much trouble in any colony. It is one that has been used successfully in many.

SELECTION COMBINED WITH INBREEDING. It is assumed that, in any species and with any degree of inbreeding, constant selection will be maintained to eliminate sibships in which undesirable mutations appear or which are reduced in number by lethal genes operating at early stages of development. Selection may also be practised to head the strain in any direction particularly desired for some special type of work. The available evidence indicates that it should not be difficult to make any strain more resistant to some particular disease, or less so, if that were desired, to have it with an unusually high or low requirement of some vitamin, to make its members capable of living longer than their unselected fellows, or to modify its sensitivity to various environmental influences.

To some extent, this has already been done in the differentiation of breeds and varieties, which, as was earlier pointed out, represent varying degree of inbreeding and hence of uniformity. The fact that these may differ in important physiological traits does not imply conscious selection by the founders of the breeds, but, more likely, a natural selection by widely different environments in different parts of the world. Anyone doing biological research should take advantage of these breed differences, in so far as they may facilitate certain types of work, and beware of them if they are likely to confuse others. For example, in addition to the differences between Leghorns and heavier breeds already mentioned, there are important differences in susceptibility to pullorum disease, and to the nematode, *Ascaridia lineata,* in ability to withstand high temperature, in thickness of egg shell, in broodiness, and in response to hormones. Some of these, and possibly all, are quite independent of differences in body size. Similarly, as Casey *et al.* (1936)[27] have shown, breeds of rabbits differ significantly in the number and proportion of different kinds of blood cells.

RECOMMENDATIONS. To reduce variability, therefore, the biologist should (1) learn the range of variation to be found in the species with which he works, (2) select the most suitable breed or variety for his purpose, (3) should, if possible, obtain foundation stock of a strain or strains already differentiated and proven suitable for that purpose, and preferably somewhat inbred, and (4) maintain the colony without in-

[27] Casey, A. E., F. D. Rosahn, C. K. Hu, & L. Pearce. J. Exper. Med. 64: 453–469. 1936.

troducing new blood and with as much inbreeding as can be practised without endangering vigor and reproduction. As new mutations appear, it will be desirable to learn, either by direct experimentation or from friendly geneticists, how they are inherited (and, hence, how best eliminated if that be necessary) and also to what extent they may affect the work in progress. The example of the yellow mouse was used deliberately to show that even an apparently harmless simple mutation affecting hair color may have manifold far-reaching effects.

VARIATION IN ANIMALS RECENTLY DOMESTICATED

In the short period of seventy-six years that has passed since Darwin first drew to the attention of biologists the great extent to which animals vary under domestication, that variation has undoubtedly been somewhat increased. To some extent, the very animal colonies under consideration in this symposium are responsible for the differentiation of races, varieties, or even breeds There are over 200 breeds or varieties of the domestic fowl, in many colors and patterns, and varying in size from a 500-gram bantam to a Jersey Giant ten times as heavy. In striking contrast, among over 90 skins of *Gallus gallus* (which is generally accepted as the ancestor of all domestic fowls), taken in the wild from Lombok through the Dutch East Indies, through Burma and Indo-China to Hainan, and now available at the American Museum of Natural History in the Whitney and Rothschild collections, there is so little variation in size and color that the geneticist accustomed to the motley assortment in the domesticated population, but not to the subtle intricacies of taxonomy, finds it difficult to see the differences between the four sub-species of the Red Jungle Fowl that are now recognized by the "splitting" ornithologists.

Similarly, it is to be expected that any species recently domesticated will, at first, show remarkable uniformity in size, in color, and probably also in physiological traits. So it is with the golden hamster, *Cricetus auratus,* which is currently becoming popular as a laboratory animal. However, as more and larger colonies are established, mutations will be accumulated and races will be differentiated, as has already been done with mice, rabbits and other rodents. This is especially so if the recently domesticated species is one that appeals to the fanciers, as has the budgerigar, *Melopsittacus undulatus,* or if it produces mutations of great economic value, as have the mink and the fox.

The remarkable uniformity of a wild population does not mean genetic homozygosity except for the genes that have been subject to nat-

ural selection. Lethal genes carried in the heterozygous condition are sure to be present, and, for that reason, close inbreeding is a more risky procedure with a newly domesticated species than with one long domesticated and in which some degree of inbreeding may have already occurred.

GENES, ENVIRONMENT AND VARIABILITY

Variation is inevitable, even in a stock so highly inbred that it is almost a pure line. Whether it be the fruit flies that develop abnormal abdomens on moist food but not on dry ones, the turkeys that get pendulous crops in the hot climate at Davis, California, but not at Tomales, which is closer to the sea (Asmundson and Hinshaw, 1938),[28] or the rats that develop avitaminosis on one diet but not on another; these are only illustrations of the fact recognized by all geneticists that the phenotype is the end result of the interaction of the genotype with the environment. Sometimes, the environmental influence is greatest when it is quite unsuspected. It has recently been shown (Hutt, *et al.*, 1944)[29] that mortality in adult fowls from neoplasms, principally lymphomatosis, depends to a great extent upon the exposure to the causative agents during the first two weeks of life. The use of two brooder houses, apparently alike in most respects, may result in quite different mortality rates from this disease during a test period that doesn't begin until 21 weeks after the chicks have left those brooder houses. (TABLE 1.)

TABLE 1.
DEATHS FROM NEOPLASMS BETWEEN 160 AND 500 DAYS OF AGE, PER CENT.
DATA FOR THREE YEARS

Strain	Birds number	Chicks brooded for the first two weeks in:	
		House F.	House B.
C. Resistant	1604	14	5
K. Resistant	942	15	6
Susceptible	768	30	16

The fact that this difference was quite unsuspected during six years of selection did not prevent the differentiation of the resistant and susceptible strains, but it did make that process much slower than it would have been had it been recognized earlier that exposure is more severe in a brooder house 40 feet from adult fowls than in one 40 yards away.

[28] Asmundson, V. S., & W. E. Hinshaw. Poultry Sci. 17: 276–285. 1938.
[29] Hutt, F. B., R. K. Cole, M. Ball, J. H. Bruckner, & R. F. Ball. Poultry Sci. 23: 396–404. 1944.

Finally, lest the impossible be expected, it should be emphasized that there are some genetic characters in which variability cannot be eliminated even in highly inbred stocks and in a constant environment. The mutation, *radius incompletus,* is a simple recessive variant of wing venation in *Drosophila funebris.* It is not influenced by the environment, but its penetrance, or the proportion of homozygotes that show the character, varies greatly in different lines so highly inbred as to constitute practically pure lines. Timofeeff-Ressovsky (1927)[30] found that, in one of these, 39 per cent of the homozygous recessives failed to show the character, although breeding tests proved that those not showing it were identical in genotype with those that did. In other lines, the penetrance was higher, the differences being attributable to the action of other genes in the different lines. Other cases of incomplete penetrance are known.

The white spotting of guinea pigs depends upon a pair of recessive genes with major effects. Many genes with minor effects apparently permit different degrees of spotting in different lines, but Wright (1936)[31] found that even in highly inbred isogenic strains, the amount of white could vary from a mere trace to 100 per cent white in strains in which the median grade was about 50 per cent. By far the larger proportion of this variability was attributed by him to "developmental accidents" not genetic in origin. Obviously, inbreeding can not eliminate all variation in all characters.

IN DEFENSE OF GENETIC VARIABILITY

There can be no doubt that, for most kinds of laboratory work, it is desirable to have animals in which genetic variability is reduced to the lowest possible minimum. The nutritionist assaying vitamin D in fish oils is not concerned with the genetic variations in the rate of ossification in the chicks he uses as test animals but with differences in the potencies of different oils. Similarly, the cancer specialist, studying the role of environment in the causation of that disease, wants a minimum of variability in his resistant and susceptible strains. However, these strains could not have been provided for his use unless there had been genes responsible for differences in susceptibility and resistance by the accumulation of which the highly resistant or susceptible strains were developed.

Differences between individuals and between strains may lead to further research to explain discrepancies, and consequently to more

[30] Timofeeff-Ressovsky, N. W. Genetics 12: 128–198. 1927.
[31] Wright, S. Genetics 21: 758–787. 1936.

complete knowledge and broader applications than are likely when uniform results are obtained in the first experiment.

For example, perosis, or slipped tendon, is a genetic and nutritional abnormality of growing chicks in which, because of faulty osteogenesis, the tibio-tarsal joint becomes enlarged and twisted so that the tendon from the gastrocnemius muscle frequently slips out of the intercondylar groove, thus causing the chick to be crippled. It was shown by Wilgus, Norris and Heuser (1937)[32] that this condition was prevented, in most cases, by adding manganese to the diet. However, it was later found (Gallup and Norris, 1939)[19] that, while 30 parts per million prevented perosis in all Leghorns and 50 p.p.m. did so in most New Hampshires, some chickens developed the abnormality even on diets containing twice the latter amount. These exceptions led to further research and the discovery that some birds need not only more manganese than their fellows for normal osteogenesis, but also more choline (Jukes, 1940)[33] and that all require biotin (Jukes and Bird, 1942).[34] Perosis is rare in Leghorns kept on normal diets. If the heavier breeds were not comparatively susceptible, the abnormality would not have received as much study as it has in the past 15 years. That research, in turn, might have stopped with the discovery that some birds need more manganese than others, were it not for few non-conformists whose special requirements included also an excess of choline.

Although genetic variability has undoubtedly been most useful in providing the variants from which superior strains of animals and plants have been bred by selection, it should not be difficult to find, in nearly any field of biology, cases like the example just given, in which a lack of genetic purity has led directly to a valuable extension of knowledge. In such cases, genetic variability is an unrecognized blessing in temporary disguise.

SUMMARY

Litter mates differing by a simple unifactorial mutation may be extremely different in form, in physiology and in viability. Multiple factors with important cumulative effects cause important differences between breeds, varieties and strains with respect to nutritional requirements, resistance to disease, and other important functions.

To reduce genetic variability, the laboratory worker is recommended to select from the existing breeds and strains those already proven suit-

[32] Wilgus, H. S., Jr., L. C. Norris, & G. F. Heuser. J Nutrit 14: 155–167. 1937.
[33] Jukes, T. H. J. Biol Chem. 134: 789–790. 1940.
[34] Jukes, T. H., & F. H. Bird. Proc. Soc. Exper. Biol. Med. 49: 231–232. 1942.

able for his purpose, and to reduce heterozygosity still further by the most intense form of inbreeding that is compatible with vigor and good reproduction. Results to be expected with different degrees of inbreeding and with selection are discussed.

It is pointed out that variation is comparatively slight in animals recently domesticated, that some variability may persist even in highly homozygous lines, and that a little variation in animal colonies is sometimes responsible for the extension of knowledge.

DISCUSSION OF THE PAPER

Dr. L. S. Strong (*Yale University School of Medicine, New Haven, Conn.*):

First, I want to congratulate Doctor Hutt for his excellent paper, presented at this symposium.

On two points, I disagree with Doctor Hutt. I would hesitate to use all variable strains of all animals. The investigator is used to his own animals. In other words, we are beginning to report results of two heterozygous resultant quantitative variables to demonstrate results in the field of cancer research. Little variation is gotten in any of them. I am quite sure it will probably not take place in our generation. Almost every variation that I have is not qualitative, but is quantitative. Age is a more important variable with quantitative strains of animals and should be taken into consideration.

Dr. T. F. Zucker (*College of Physicians and Surgeons, Columbia University, New York, N. Y.*):

From what Doctor Hutt has stated and the answers he has given to the questions, it might be helpful to mention the work done with rats. I have been looking for a geneticist working with rats and Doctor Gowen and I would like to know where geneticists are now working with them.

Doctor Hutt has spoken about Doctor Palmer's work with rats. One feature of this was production by selection of a strain in which more production was maintained. During the course of our work (rats and resistance to rickets, etc.), both strains were lost and no animals retained. Our animals gradually died off. Both strains presented a kind of problem which must be discussed.

An important quantitative problem is body size. Doctor Mendel pointed out stature as genetic and rate of growth as due to nutrition. In most cases, you produce mice by selection. Body weight and intrinsic weight are extreme. Food requirement is regulated by body weight. There is extreme variability in the rat. Doctor King has done important work on body weight by use of the coefficients of variability in body weight at given ages of the animals and Doctor Smith's work with inbred strains should be mentioned.

In our work at Columbia with the rat, the weight remained constant in all different types, as soon as we began selection for body size. In about three generations, we raised the weight from 45 to 60 in males and from a little over 50 to somewhat under 70. I wonder whether there is more information available on the various features in the genetics of the rat?

Reply by Dr. Hutt:

Till recently, most geneticists working with rats were more concerned with those variations brought about by genes which affect color, etc., than with the physiological characters, but studies of the latter are now being made. Some selection has been made in respect to requirements of vitamin D.

Dr. Myron Gordon (*American Museum of Natural History, New York, N. Y.*):

The National Research Council maintains a Committee of which Doctor Landauer is Chairman and includes Doctor Little of Bar Harbor. It is responsible for information concerning poultry I am in charge of fishes.

I want to point out that some sort of committee should be established which will be able to supply information concerning the availability and location of various pure strains of animals needed for laboratory experimentation.

Capt. Roy Nichols (*Army Medical and Veterinary Schools, Washington, D. C.*):

It is not possible to call animals truly genetically identical in the genetic sense, until we have overcome the selection problem in the different laboratories and have standardized the strain genetically. Cannot each laboratory standardize its strains? Inevitably, when one strain has been kept genetically identical for several generations by selection and their location is changed, variation will occur in the animals.

Reply by Dr. Hutt:

A difference in the environment in different laboratories could bring about a difference in the strains. If one laboratory maintains its thermostat at 68° and another one at 80°, it would not be surprising if after several years the strains differed in their temperature requirements.

THE MATING OF MAMMALS

By Carl G. Hartman

Department of Zoology and Physiology, University of Illinois, Urbana, Ill.

DEFINITION

The term, mating, may be used transitively or intransitively. In the latter sense, mating is usually used synonymously with copulation. More rarely, and more or less anthropomorphically, where the association of male and female is for a season or "for life" (foxes?, swans?, elephants?), the word, mating, is applied to this more permanent relation. The term in this sense begins to have social significance, and sex to become a factor in the organization and stability of the family.

In a state of nature, in herds on the range or large mixed colonies of laboratory mammals, the animals are free to choose their mates. But breeders or laboratory experimenters usually wish to control the mating of their animals. Hence, mating is employed in the sense of placing together males and females which are at other times isolated. In this transitive sense, the word is used synonymously with "breeding" or "hand mating," terms employed by breeders of the larger farm animals.

The present paper is concerned with such controlled matings as are carried out for exactitude in experimental work, or to increase fecundity by insuring fertilization of the ova.

The controlled breeding of animals by man antedates recorded history. It is only recently that it has been gradually emerging from the empirical to the scientific and becoming a science rather than an art. This advance is due to the great strides made, chiefly since the turn of the century, in the study of the physiology of reproduction. Since the mating techniques are based on scientific principles, some of the principles will be outlined before individual species are considered. After that, the common laboratory species will be discussed, together with some of the larger animals of agricultural importance.

RUT

Rut, or "heat," or estrus, is the limited period in which the female accepts the male. Every species has its own characteristic behavioral manifestations of estrus (see Young, 1941,[1] for a summary of these). Thus, the cow in heat will mount other cows; the rat will dart, vibrate her ears, and exhibit lordosis when approached by other rats. In all

[1] Young, W. C. Quart. Rev. Biol. 16: 135–156, 311–335. 1941.

instances, the female in heat becomes very active and restless, and her food intake is greatly lessened. Teleologically speaking, this activity is favorable for the female to seek out the male.

In a broad sense, heat is closely related to ovulation, except in pathological cases, as, for example, in nymphomaniac cows. This holds rather strictly for the lower mammals, but is less apparent in monkeys and apes, in which we see the dawn of the social significance of sex. In man, finally, the picture is almost totally obscured, for a rise in sex desire is seldom reported as occurring in the middle of the cycle when ovulation certainly takes place (Ball and Hartman, 1935[2]; Hartman, 1936[3]).

THE PHYSIOLOGICAL BASIS OF ESTRUS

It has been known from time immemorial that removal of the gonads results in almost complete blotting out of the sexual response in the male (eunuchs, geldings, steers, capons). After the advent of abdominal surgery, the same was found true of the female also. It remained for the modern endocrinologist and the behaviorist to analyze the internal chemical milieu necessary to sensitize the nervous system so that the adequate stimulus (presence of the mate or sex partner) may elicit the appropriate response—i e., "male" behavior of the male in the presence of the female, "female" behavior of the female in the presence of the male. It has been shown, however, in appropriate sex hormone experiments, that the chromosomal constitution (if that conditions sex behavior) may be overridden by the appropriate heterosexual hormone (Stone, 1939[4]; Ball, 1937,[24] 1940[20]; Beach, 1942[4a]). Certain forms of behavior, as mounting by a female in heat (cow, guinea pig, rat), have been interpreted as "male"; therefore, for the female, "homosexual" behavior. But this seems absurd, since it is a quite common form of behavior of females when in heat, in fact, is advisedly regarded by Young (1941)[1] as a helpful indicator of heat in the guinea pig, as it is also in the cow. The matter is of more than academic interest to the rabbit breeder, in that such mounting behavior of an estrous female often results in one or the other or both partners ovulating and becoming pseudopregnant and therefore sterile for the succeeding 16 days.

Sex behavior in the male is normally brought about by the male hormones, testosterone and androsterone. In the female, estrone and es-

[2] Josephine, & C. G. Hartman. Am. J. Obstet. and Gynec. 29: 117–119. 193[illegible].

[3] Hartman, Carl G. Time of Ovulation in Woman. Williams & Wilkins Co. Baltimore. 1936.

[4] Stone, Calvin P. Chapter on Sex and Internal Secretions. Edgar Allen, Ed. Baltimore. 1932 (1st Ed). 1939 (2nd Ed.).

[4a] Beach, F. A. Psychosomatic Med. 4: 173–198. 1942.

tradiol, secreted by the ovarian follicles, condition female sex behavior. In the intact female, as estrus approaches, her sex behavior becomes more and more pronounced, in correlation with the growth of the graafian follicle, and the most intense manifestation synchronizes with the maximal distension of the follicle. "Silent" estrus, that is, morphological changes leading to ovulation without overt behavioral manifestation, also occurs.

In the castrate, estrogens alone are able to call forth mating behavior. Indeed, I have seen an ovariectomized bitch kept in heat constantly for several years. In the castrated rat and guinea pig, however, not all treated females will come into heat, but the refractory individuals, after being primed with estrogens, can be brought promptly into heat by a dose of progesterone. Furthermore, progesterone, added to estrogen, greatly intensifies the response. It seems very probable that, in the normal cycle, the ripe graafian follicle produces progesterone, which precipitates the copulatory response (and ovulation!—Everett, 1944),[5] for there is collateral evidence of progesterone production by the unruptured follicle (see review article by W. C. Young, 1941).[1]

Upon the establishment of the corpus luteum and the initiation of the "progesterone" phase of the cycle, the female becomes not only indifferent to the male but, in some species, positively antagonistic.[5a] In the latter case, the sexes should be separated after mating, at least not kept together in close quarters, for the female is likely to injure the male. I have seen a small opossum female, during the night after her heat period, kill a male double her weight.

To summarize: estrus in the lower mammals is limited to a few hours or a few days of the sexual cycle. The "cycle" may comprise a whole year in "monoestrous" animals—those having a single rutting season—such as deer, many wild rodents, and probably all marine mammals. The dog comes into heat twice, the cat perhaps six times in a year, the long period between estri being known as the "anoestrum" (Heape), in contrast to the "diestrum," the interval between short, regularly recurring estri of "polyoestrous" mammals.' Some mammals—as, forexample, monkeys—experience cycles the year round, but are fertile only in a portion of the year (Hartman, 1931;[6] see also Zuckerman, 1931[7]). In some of our highly domesticated mammals (rat, guinea

[5] **Everett, J. W.** Endocrin. **34:** 136–137. 1944.
[5a] For the pathological human counterpart, see **Benedek, Th., & B. B. Rubenstein** Psychosomat. Med., Mono. **III.** 1942.
[6] **Hartman, Carl G.** J. Morph. **12:** 129–140. 1931.
[7] **Zuckerman, S.** Proc. Zool. Soc. London. **25B:** 343. 1931.

pig, rabbit, cow), the non-breeding season has all but been eliminated; less so in the horse, and still less so in the sheep, goat and ferret; the last three still experiencing rather marked though variable anestrous periods.

THE TIME OF OVULATION

In animal breeding, it is of the utmost practical importance to know exactly when the egg or eggs leave the ovary. It is, of course, true that multiple matings tend to insure fecundity to the maximum; but with expensive animals like stallions and bulls or even rams and boars, the animal husbandman must economize by making his males serve as many females as possible. The embryologist or the physiologist in the laboratory, less concerned about the cost of maintaining a few extra males of the laboratory rodents, finds great advantage in the precision with which the hour of ovulation may now be predicted in several species. Once the hour or even the day of ovulation is known, the female may be served at the most fertile moment. The necessity for precision in this matter arises out of the low period of viability of sperms in the female genital tract (about 30 hours) and the still more critical period of fertilizability of the ovum (about 10–15 hours), as is now generally recognized (Hartman, 1924,[8] 1932,[9] 1939;[10] Blandau and Young, 1939;[11] Blandau and Jordan, 1941[12]).

Knowledge of the time of ovulation with reference to outward signs —in the lower mammals, the beginning and end of estrus, in the primates, menstruation—is gradually being built up. Very accurate data, as will be seen below, exist for rat, mouse, guinea pig, rabbit, ferret, and cow, less accurate for sheep, goat and mare.

THE ENDOCRINOLOGY OF OVULATION

Categorically stated, the growth of the graafian follicle is brought about by the follicle stimulating hormone (F.S.H.) of the anterior pituitary, the rupture of the follicle by the addition of the luteinizing hormone (L.H.). It is, therefore, possible to precipitate ovulation and thus fix the time of ovulation with certainty. The classic example is the estrous rabbit, which will ovulate about 10 hours after injection of any of the ordinary gonadotrophs (Friedman test for pregnancy). The estrous cat and certain hibernating bats are further examples. Mirs-

[8] Hartman, Carl G. Am. J. Obst. & Gynec. 7: 40–43. 1924.
[9] Hartman, Carl G. Contributions to Embryology 13: 1–162. 1932.
[10] Hartman, Carl G. Chapter on Sex and Internal Secretions. Edgar Allen, Ed. Baltimore. 1939.
[11] Blandau, R. J., & W. C. Young. Am. J. Anat. 46: 303–329. 1939.
[12] Blandau, R. J., & E. S. Jordan. Am. J. Anat. 68: 275–291. 1941.

kaia and Petropavlosvsky (1937)[13] have taken advantage of the technique to precipitate ovulation in the estrous mare and hence to be certain of ovulation 30–48 hours after the treatment (1000 M. U. of human chorionic hormone).

For reasons which the writer has stated elsewhere (1942), attempts to (a) cure sterility, (b) increase litter size, and (c) interpolate an additional pregnancy in the anestrous period of animals by means of the administration of gonadotrophic hormones are still in the experimental stages. It is true that astounding results have been achieved in spots; but the overall fecundity of a flock or herd or colony has not been increased and the expense involved has been great.

The chief loss in cases of induced ovulation is due to the failure of the treated females to mate. "Silent" heat (ovulation without overt signs of estrus) has been reported for numerous species and may be expected in all. The phenomenon is most likely to occur at the beginning and at the end of the breeding season. "Silent" heat has been discovered in horses and cows in the course of routine palpation of the ovaries; in sheep, guinea pigs and rats, in the course of autopsies. Ovulation without estrus represents ova wasted, and there is nothing the animal breeder can do about it at present.

SPONTANEOUS VS. NERVE-INDUCED OVULATION

It has been known for a hundred years (ever since Barry and Bischoff in the 1840's) that the rabbit normally ovulates only after copulation, although, as late as 1908, the subject formed an amusing controversy between Ancel and Villemin, on the one hand, Dubreuil and Regaud, on the other (C. R. Soc. Biol. Paris, Vol. LXV), in which the former did the arguing, the latter the experimenting. Villemin even asks whether one is to believe that a rabbit female would ovulate by being ogled by a male in a nearby cage!

The mechanism by which nerve impulses are set up by the act of copulation is now clear in its broad outlines (see review in Allen, 1939[14]). Through nerve channels, via the hypothalamus and the hypophyseal stalk, impulses from the erogenous zones reach the anterior pituitary and there cause the outpouring of the ovulatory hormones. As a result, ovulation occurs some time later (within about 10 hours in the rabbit, 35 in the ferret).

In the class with the rabbit, which does not ovulate spontaneously,

[13] **Mirskaia, L. M., & V. V. Petropavlovsky.** Prob. An. Husb.: 22–39. 1937. (Abst. in Am. Breed. Abstr. 5, Dec. 1937.)

[14] **Allen, Edgar.** Sex and Internal Secretions: 636–641. Baltimore. 1939.

are the cat, the ferret and the marten; and to these O. P. Pearson (1944)[15] has added the remarkable case of the short-tailed shrew, in which the female requires the stimulus of a half dozen copulations daily for several days to consummate ovulation.

ARTIFICIAL INSEMINATION

Under the leadership of Soviet scientists, the techniques of artificial insemination were brought to a high degree of efficiency and are now widely used in all countries. Males are conditioned to deposit their semen in artificial vaginae and this is then diluted, so that a single ejaculation serves to inseminate a variable number of females, depending on the species. By this method, for example, one ram, in a single season, is said to have sired over 2000 lambs.

This topic hardly comes under the subject under discussion. However, it should be noted that the males are conditioned to "mate" with dummies, and soon do so in preference to mating with females in heat.

Artificial insemination has a place not only in animal husbandry, but also for experimental purposes in the laboratory (cf. Blandau and Young,[11] Blandau and Jordan[12]).

CHANGING THE DIURNAL CYCLE BY MEANS OF LIGHT

Nocturnal animals mate at an hour of the night which greatly inconveniences the laboratory worker. It has been shown by Hemmingsen and Krarup (1937)[16] that, by turning night into day and day into night by means of electric lighting, the time when the majority of females come into heat may be changed to suit the experimenter. This is now very generally practised. It is recommended that the colony be illuminated from 6:00 p m. to 6:00 a m , and be darkened from 6:00 a m. to 6:00 p m.

VARIABILITY OF PHYSIOLOGICAL EVENTS

What is "normal" in physiological processes? Perhaps variability is the normal thing. If one examines large numbers of data, such as those of Long and Evans (1922)[17] on the rat, Bacsich and Wyburn (1940)[18] on the guinea pig, Andrews and McKenzie (1941)[19] on the

[15] Pearson, O. P. Am. J. Anat 75: 39–93. 1944.
[16] Hemmingsen, A. M., & N. B. Krarup. Kgl. Danske Vidensk. Selskab. Biol. Med. 13: 1–61. 1937.
[17] Long, J. A., & H. M. Evans. Mem. Univ. Calif. 6: 1–148. 1922
[18] Bacsich, P., & G. M. Wyburn. Proc. Roy. Soc. Edinb. 60 (Pt. I): 33–39. 1940.
[19] Andrews, F. N., & F. F. McKenzie. Res. Bull. 329, Agric. Exp. Sta. Univ. of Mo. May. 1941.

horse, one is struck with the wide variations in both the total length of the estrous cycle of animals and in the segments of the cycle. Yet, within wide limits, judged by reproductive performance, variations are "normal."

It follows from this that, in fixing mating time, one must allow for some variation; for example, in the time of ovulation with reference to some other event of the cycle. This variability will detract somewhat from the precision of our mating procedures, but this is inevitable.

TECHNIQUES OF THE ANIMAL BREEDER

In the following summary, some details concerning special techniques in the control of timed matings will be discussed, but only where either exact data on the time of conception is desired, as for securing developmental stages of the embryo, or where economic considerations make it desirable or even imperative that insemination should result in pregnancy in a higher and higher percentage of cases, as in the breeding of high-bred horses and cattle.

Much animal breeding is done still, as always, by simply allowing the sexes to mingle at will and mate when a given female is in heat. As in the past, sheep and goats, horses and cattle are still self-propagated, as it were, on the range, where there are still wide open spaces and where man-power is scarce. In many laboratories, animals are reared in mixed colonies when numbers of adult animals are needed without regard to genetic constitution or exact age of the individuals. Thus, guinea pigs seem to thrive in sizable groups in pens, on the floor in space set aside for them in a convenient corner. Mice, rats, hamsters, kept—males and females together—in large cages, leave the experimenters the maximum of offspring with a minimum of effort.

The following remarks will be restricted to some of the principles which may serve as guides in the "hand breeding" of the various species. For symptoms or overt signs of heat in these, the reader is referred to the comprehensive review of W. C. Young (1941).[1]

THE RAT

Ovulation in the rat is related, temporally, with several easily determinable signs, namely, the vaginal smear, the voluntary or "spontaneous" activity, and the sexual response or estrus. A less exact procedure is by daily examination of females caged with males for the vaginal plug or the coagulated semen of the male.

The last method is sufficiently exact for most purposes, when it is recalled that most females come into heat between 6:00 p.m. and 3:00 a.m., with the peak round 11:00 p.m. If one institutes a regular 12-hour artificial day by electric illumination, one can work out for his own colony a curve of hour of ovulation and make his calculations on the basis of finding vaginal plugs early in the morning. Occasionally, a plug will fall out, giving a false negative, which, however, may be avoided, if one takes a vaginal smear in suspicious cases, as, for example, where one finds a swollen vulva and no plug.

In this connection, the novice, in the handling of a rat colony (the same holds also for mouse and probably hamster colonies), should have called to his attention, first, that copulation and ejaculation are not synonymous, for the rat may copulate many times without ejaculating. Second, at least two plugs should be allowed before the male is removed from the cage, for Ball (1940)[20] found that one-plug matings frequently afforded insufficient stimulus to the pituitary to activate the corpora lutea for progesterone secretion, indispensable for pregnancy changes in the uterus.

The original data of Long and Evans (1922)[17] on the relation of the vaginal smear picture to the time of ovulation have served as a good guide for many years. A female rat mated when the smear consists of large epithelial cells only, or with the addition of a few scales (early cornified stage), is likely to become pregnant. Young, Boling and Blandau (1941)[21] have reinvestigated this schedule and find that, by the time cornification is completed, the female rat has already ovulated.

Taking advantage of the voluntary running activity cycle of the female rat (Wang, 1923)[22] Farris (1942)[23] has determined quite precisely that sexual receptivity follows soon after her running begins to increase markedly and that ovulation occurs a little over 10 hours later in the non-mated, 8 hours later in mated individuals. Females mated while the activity is on the increase become pregnant in 9 out of 10 cases. The number of females that can be bred according to this schedule by this method is, of course, limited by the number of individual activity cages at one's disposal.

More accurate timing of ovulation and fertilization of the ovum is afforded by the method of determining the exact beginning of behavioral estrus. One does not need a male to make the test for

---

[20] Ball, Josephine. Am. J. Physiol. 130: 471–474. 1940.
[21] Young, W. C., J. L. Boling, & R. J. Blandau. Anat. Rec. 80: 37–45. 1941.
[22] Wang, G. H. Comp. Psych. Monogr. 2: 1–27. 1923.
[23] Farris, E. J. Anat. Rec. 84: 4. 1942.

Ball (1937)[24] showed that the estrous female responds to the "finger test." This consists of quick, rhythmic clasping of the female just in front of the iliac crests, whereupon the rat, if she is in estrus, responds with lordosis, as when clasped by the male. Blandau, Boling and Young (1941)[25] have described and illustrated a modification of the test, which is applicable to the guinea pig and the hamster also.

Ovulation occurs about 10 hours after the onset of heat (Boling, Blandau, Soderwall and Young, 1941).[26]

In the work with the rat little advantage has been taken of the post-parturitional ovulation characteristic of many rodents. According to Blandau, Jordan and Soderwall (1940,)[27] the average interval between parturition and the beginning of heat was 18.5 hours (range 4–36) and that the temporal relation between the beginning of heat and ovulation was the same as in nonpregnant rats.

THE MOUSE

Certain facts, set forth above for the rat, hold also for the mouse. Ovulation occurs at a time when the vaginal contents include types of cells corresponding to those in the rat (Allen, 1922,[28] Snell *et al.*, 1940).[29] The latter authors report ovulation occurring in their mice shortly after midnight and not over 2 to 3 hours after the beginning of heat.

With those data in mind, to secure timed material, the breeder would do well routinely to look for vaginal plugs in the morning and to calculate the most probable hour of fertilization according to Snell and co-workers. It should be mentioned, however, that these authors used artificial "day and night" so as to control the rhythm of the mice more accurately; however, Lewis and Wright were able to secure many mouse ova in cleavage, without special illumination of the colony, by looking for vaginal plugs each morning in females caged with males.

Sobotta (1895)[30] secured nearly 1500 mouse ova by utilizing the postpartum ovulation, but he made no generalization as to the exact time when ovulation occurred. Long and Mark (1911)[31] studied this point carefully and reported that, in their mice, ovulation occurred 14½ to 28½ hours after delivery of the young.

[24] **Ball, Josephine.** J. Comp. Psych. **24**: 135–144. 1937.
[25] **Blandau, R. J., J. L. Boling, & W. C. Young.** Anat. Rec. **79**: 453–463. 1941
[26] **Boling, J. L., R. J. Blandau, A. L. Soderwall, & W. C. Young.** Anat. Rec. **79**: 313–331. 1941.
[27] **Blandau, R. J., E. S. Jordon, & A. L. Soderwall.** Anat. Rec **78**: 58. 1940.
[28] **Allen, Edgar.** Am. J. Anat. **30**: 297–371. 1922.
[29] **Snell, G. D., E. FeKete, K. P. Hummel, & L. W. Law.** Anat. Rec. **76**: 39–54. 1940.
[30] **Sobotta, J.** Arch. f. Mikr. Anat. **45**: 15. 1895.
[31] **Long, J. R., & E. L. Mark.** Carnegie Inst. Wash. Publ. No. **142**. 1911.

Because of the smallness of the mice, they may be bred in numbers in relatively small cages, litters being removed and labeled as they appear.

THE GUINEA PIG

Nearly a century ago (1852), Bischoff[31a] and, after him, Hensen (1876)[31b] and Rein (1883)[31c] utilized the postpartum ovulation in order to mate their guinea pig females to secure "timed" embryological material. The method is still useful.

Working with the guinea pig, Stockard and Papanicolaou (1917)[32] gave us the "vaginal smear" method by which estrus and ovulation could be timed with a fair degree of accuracy. It has since been shown, however, by Young and co-workers (Young, 1937)[33] that, as in the rat, the time of ovulation may be predicted with a high decree of accuracy, by noting the time of beginning of estrus and calculating 10 hours forward. By the time the fully cornified stage is reached by the vaginal mucosa, most of the females are already out of heat. The rat and the guinea pig are alike in ovulating about 10 hours after the beginning of heat. The rabbit ovulates about 10 hours after mating.

THE RABBIT

As already stated, the rabbit belongs to that group of mammals which do not ovulate spontaneously, but only in response to a stimulus—either the nervous stimulus to the pituitary as a concomitant of copulation, or in response to exogenous gonadotrophic hormones (cf. cat, ferret, mink and marten).

During the breeding season—or at least that season in which rabbits breed best (cf. Marshall, 1922;[34] Hammond, 1925;[35] Hammond and Walton, 1934[35a])—it had been supposed that the doe remained in estrus ("constant estrus") for the season, in consequence of the presence of large follicles in the ovaries, and, hence, that she could be mated at any moment. A single crop of follicles was supposed to remain intact for the season. Neither of these conclusions is correct. In the first place, even a casual experience with a rabbit colony teaches that does pass through short periods when they will not take the buck. In the second

[31a] **Bischoff, Th. L. W.** Ent -gesch. d. Meerschweinchenelesꞏ Giessen 1852
[31b] **Hensen, V.** Zeitschr. f. Anat. u. Entwickl.-gesch: 213, 351. 1876
[31c] **Rein, G.** Arch. f. mikr. Anat. **22**: 233. 1883.
[32] **Stockard, C. R., & G. Papanicolaou.** Am. J. Anat **22**: 225–283. 1917.
[33] **Young, W. C.** Anat. Rec. **67**: 305–325. 1937.
[34] **Marshall, F. H. A.** The Physiology of Reproduction, London. 1922.
[35] **Hammond, J.** Reproduction in the Rabbit. London. 1925.
[35a] **Hammond, J., & A. Walton.** J. Exp. Biol. **11**: 307–319. 1934.

place, it is now known that follicles come and go in cycles of about 15 or 16 days, one set degenerating while another set comes on (Smelser, Walton and Whetham, 1934).[36] It is probably in the transitional period, while the new set of follicles is growing and the old set is retrogressing, that the doe lacks interest in the male.

Because of the sensitivity of the ovulatory mechanisms, it happens that, as stated above, females may stimulate each other to the point of ovulation! Does should, therefore, be kept isolated if "estrus" individuals are desired. Advantage may also be taken of the postpartum ovulation in the rabbit (Weil, 1873).[37] The postpartum doe constitutes, indeed, the most reliable stage for the Friedman pregnancy test. The government bulletins by Templeton (1940)[37a] and Templeton, Ashbrook and Kellogg (1942)[37b] will prove of practical use to every rabbit breeder.

THE GOLDEN HAMSTER

Too little is known about this species, which is destined to become an important laboratory rodent. The estrous cycle is about 4 days in length. We have seen hamsters mate in the early evening but have not studied them in detail. They are as easy to raise as rats. We have followed the practice of leaving one male with several females and isolating a female as soon as she is palpably pregnant. The hamster, like the rat, should be handled gently with frequency to keep it tame and tractable.

THE OPOSSUM

The vaginal smear of the opossum follows the general picture of the rat and the guinea pig in that the cornified stage is associated temporally with ovulation (Hartman, 1923).[38] There seems to be a considerable "postoestrous" period (between estrus and ovulation) not yet determined with any accuracy.

Before the advent of the vaginal smear technique, the writer was able to determine the approach of estrus in the opossum by the increasing swelling (edema) of the mammary glands (Hartman, 1921).[39]

Timed matings are best made in either of two ways: the less accurate one of examining the vaginal lavage mornings and evenings for

---

[36] **Smelser, G. K., A. Walton, & E. O. Whetham.** J. Exp. Biol. **11**: 352–363. 1934.
[37] **Weil.** Wiener Med. Jahrb 1873.
[37a] **Templeton, G. S.** Wildlife Circular. U. S. Fish and Wildlife Service. 31 pp.
[37b] **Templeton, G. S., F. G. Ashbrook, & C. L. Kellogg.** Conserv. Bull. **25**, U. S. Fish and Wildlife Service. 63 pp.
[38] **Hartman, Carl G.** Am. J. Anat. **32**: 353–421. 1923.
[39] **Hartman, Carl G.** Am. J. Physiol. **55**: 308–309. 1921.

sperms; or, more accurately, keeping a group of females with a number of males running free in a large dimly illuminated room under constant observation for cases of copulation. These occurred mostly at night.

THE DOG

The dog family offers several exceptional features. The males do not possess seminal vesicles. In copulation, the male "hangs" or remains interlocked with the female. The act, in common with that of most carnivores, is long, upward of 20 minutes, which is quite the antithesis of the rabbit, the goat and indeed most mammals.

Ejaculation in the dog may be divided into three stages, as determined in part by masturbating the male, in part by observation through uterine fistula (Evans, 1933).[40] First, there is a clear secretion (from Cowper's gland?) almost devoid of sperms; second, a milky secretion containing clouds of sperms, perhaps the bulk of those furnished at a given mating; third, for most of the duration of coitus, a thin secretion of the prostate, carrying sperms in relatively small numbers.

The dog ovum (as also that of the fox) is peculiar in that the first maturation division does not take place until after it reaches the oviduct. There are, furthermore, some breeding data which point to a longer period of viability for the dog ovum than that of other mammals.

It seems anomalous that, for man's oldest domesticated animal, we cannot yet state definitely when ovulation occurs with reference (a) to the beginning of sexual receptivity, and (b) to the uterine bleeding. Evans and Cole (1931),[41] in their monograph on reproduction in the dog, state definitely that the bitch ovulates "usually within a day after the first acceptance of coitus." The autopsies recorded bear out this conclusion. However, practical dog breeder F. L. Whitney (1927,[42] 1942,[43]) who has bred thousands of dogs, places the "mode" for ovulation day on the 5th day of estrus. Whitney (1942)[43] also approached the problem by double matings with males of different breeds, one early in estrus, the other later. Matings on days 1 and 2 were sterile, later matings (days 5, 6 and 7) were fertile, as judged by the phenotype of the offspring.

Artificial insemination, with or without the use of gonadotrophic hormones, has not often been resorted to in the case of the dog.

[40] Evans, Everett. Am. J Physiol. 105: 287 1933.
[41] Evans, H. M., & H. H. Cole. Mem. Univ. Cal 9: 57–101. 1931.
[42] Whitney, Leon F. Chase Magazine ——. 1927.
[43] Whitney, Leon F. Vet. Med. 35(3); See also 35(1): 59–60. 1940.

THE CAT

The domestic cat displays marked symptoms of estrus, as described by Bard (1939, 1940),[44] and others quoted by Young (1941).[1] The overt signs of heat constitute sufficient guide for the investigator to regulate timed matings. Ovulation in the cat requires an extraneous stimulus, such as: artificial stimulation of the cervix (Greulich, 1934),[45] injection of gonadotrophic hormones (Windle, 1939)[46] or, normally, copulation. Eggs are discharged from the ovary about 30 hours later—in 26–27 hours according to Gros (1936),[47] in 25 hours after mechanical stimulation of the cervix, according to Greulich and Dawson and Friedgood.

THE FOX

Because of the recent expansion of the silver fox breeding industry, the U. S. Fish and Wildlife Service has had some intensive studies made on reproduction in this form, chiefly by O. P. Pearson and R. K. Enders of Swarthmore College,[48, 49, 50, 51] in collaboration with C. F. Bassett,[52] of the U S Fur Animal Experiment Station at Saratoga Springs (Bassett & Leekley 1942;[52a] Bassett, Wilke & Pearson, 1943[52b]).

The data presented by Pearson and Enders indicate that vixen tend to ovulate on the second day of receptivity. As is the case in all of the carnivores so far studied, estrus is slow in coming on. This is reflected in the gradually increased swelling of the vulva, the most reliable outward sign. As the breeding season approaches, vixen should be examined every third day for increased swelling of the vulva. When this approaches the maximal, the female should be tested with a reliable male for acceptance, and insemination should be allowed on the day following the first acceptance.

Heretofore, because most fox males are monogamous, testing vixen for acceptance has been difficult, since polygamous males, suitable for the test, have been scarce. The writer has the impression that, in recent years, the polygamous tendency has been bred into stocks, so that more such highly desirable males are now available.

[44] **Bard, P.** Res Publ Assoc Nerv & Mental Dis **19:** 190–218 1939, **20:** 551–579 1940.
[45] **Greulich, W. W.** Anat Rec **58:** 217–224. 1934.
[46] **Windle, W. F.** Endocrin **25:** 365–371 1939.
[47] **Gros, G.** Thèse Alger. 146 pp 1936.
[48] **Pearson, O. P., & R. K. Enders.** Anat. Rec **85:** 69–83. 1943
[49] **Pearson, O. P., & R. K. Enders.** Am. Fur Breeder **17** (Jan), 32–34 1944(a)
[50] **Pearson, O. P., & R. K. Enders.** Am. Fur Breeder **17:** (July)· 24, 26, 28, (Aug)· 26, 28, 30, 42 1944(b).
[51] **Pearson, O. P., & R. K. Enders.** J. Exp Zool. **95:** 21–35 1944(c).
[52] **Pearson, O. P., & C. F. Bassett.** Anat. Rec. **89:** 455–459. 1944
[52a] **Bassett, C. F., & J. R. Leekley.** N A. Veter. **23:** 454–457 1942
[52b] **Bassett, C. F., F. Wilke, & O. P. Pearson.** Am. Fur Breeder **16:** 22–26. 1943; Black Fox Mag. **27:** 12, 13, 20, 21, 23, 25–27, 29. 1943.

FERRET, MINK AND MARTEN

Because of the value of the fur, the demand for it, and the growing scarcity of the feral population of our fur bearers, "fur farming" is steadily expanding. New mutations in mink are constantly appearing, which helps to enliven the mink-breeding industry. Marten breeding is just emerging from the experimental stage, and fisher and stoat have not been studied enough to be included in this discussion.

The Mustelidae mentioned, have, together with the most domesticated of them all, the ferret, several features in common: (1) they all require the stimulus of coitus for ovulation (Marshall, 1904,[53] 1922;[34] Pearson and Enders, Jan. 1944[40]); (2) the swelling of the vulva and dilatation of the introitus, which constitute, as in the fox, excellent signs of approaching estrus; (3) the act of copulation lasts unusually long, as in the dog.

The vaginal smear, though instructive, is less useful in the members of this family, than in the case of other mammals.

Studies on the physiology of reproduction in the mustelids have been made under the auspices of the U. S. Fish and Wildlife Service and others are in progress.

The ferret may be mated according to the vulvar enlargement. This attains, at estrus, a 50-fold increase over the anestrous condition. There is no "best" time for mating, as in spontaneously ovulating species—any time that the female accepts the male is satisfactory for breeding purposes. Ovulation occurs 30 hours, more or less, after mating. This entails the formation of the corpus luteum, in response to which the vulvar swelling recedes and pregnancy ensues—or, in case the ova are not fertilized, pseudopregnancy follows, a phenomenon which is quite marked in the carnivores.

Mink and marten are housed and bred much like the ferret (Leekley & Enders, 1941).[53a]

THE PIG

The average farmer who raises pigs for the market usually allows his sows and boars to run together and mate in their due season, nature "taking its course." The breeder of pure stock, however, must keep record of sire and dam and he therefore hand-breeds them.

From the behavior of sows, such as general activity (Altmann, 1941),[54] swelling of the vulva, behavior towards other sows (Corner, 1921),[55]

[53] **Marshall, F. H. A.** Quart. J. Micr. Sci. **48:** 323–345 1904.
[53a] **Leekley, J. R., & R. K. Enders.** Am Fur Breeder **14:** 26–28. 1941.
[54] **Altmann, Margaret.** J. Comp. Psych. **31:** 481–498 1941.
[55] **Corner, G. W.** Contrib. to Embryol. **13:** 117–146. 1921.

the onset of estrus may be roughly diagnosed, and the animals mated accordingly. Heat lasts about a day. Mating at any part of estrus seems equally effective—which greatly simplifies the matter, as in the sheep, but in contrast with the cow and the mare.

Young boars (yearlings) should not be allowed more than a single mating a day, older boars, two, three, or up to four.

THE SHEEP

Most sheep are raised on the range and breed as wild animals do, rams and ewes, juveniles and lambs all ranging together. To guide the scientific breeder, however, extensive studies have been made on reproduction of both ewe and ram: breeding season, estrous cycle, time of ovulation, breeding capacity of rams, etc. For a comprehensive survey of the literature, as well as the most detailed observations, we have the monographs of McKenzie and Terrill (1937)[56] on the ewe, and McKenzie and Berliner (1937)[57] on the ram.

There are no very definite and reliable signs of estrus in the absence of the male. Hence, reliable rams are used for testing and records kept from actual observation of the mating. Estrus lasts around a day and a half on the average, with a wide spread for variations, and the cycle is about 16 days in length. In 88 per cent of 79 cases in which the time of ovulation in relation to the beginning of estrus was accurately determined by McKenzie and Terrill, this "preovulatory" period was found to be between 24 and 36 hours.

As ovulation in the ewe occurs near the end of heat, and since the heat period is not of long duration, a single service or single artificial insemination, whenever the ewe is in heat, is likely to result in pregnancy.

Ram semen is extremely concentrated as to sperm population and admits of great dilution for purposes of artificial insemination. Thirty million sperms are used at one insemination.

There are numerous reports of artificially stimulated ovulation of ewes in the non-breeding season; but it has not yet been established that this procedure is of commercial value. (cf. Casida and Murphree, 1942).[58]

THE COW

The estrous cycle of both cow and mare is stated to be around 21

[56] **McKenzie, F. F., & C. E. Terrill.** Res Bull. **264,** Agr. Exp. Sta. Univ. of Mo. (July), 88 pp. 1937.
[57] **McKenzie, — — & V. Berliner.** Res. Bull. **265,** Agr. Exp Station, Univ. of Mo 1937.
[58] **Casida, L. E., & R. L. Murphree.** Endocrin. **31:** 545–548. 1942.

days in length. The duration of estrus is short, about 16 hours; in Zebu cattle of West Africa (Anderson, 1936),[59] less than two hours.

The cow discloses its estrous condition by marked behavior, well known to every farm boy: mounting other cows, restlessness and running, bellowing. Taking advantage of these signs and having herdsmen report the first appearance on 200 cows, Mirskaia, of the Soviet Institute of Artificial Insemination near Moscow, in 1935,[60] removed the ovaries from these animals at different periods after the onset of heat and determined from an examination of these ovaries the average time of ovulation. This proved to be one of about 36 hours. This agrees very well with the findings of Nalbandov and Casida (1942),[61] who place ovulation about 22 hours after the end of heat; and, since estrus is about 16 hours in length, we get 38 hours as the pre-ovulatory period.

One would be inclined to mate or artificially inseminate a cow as soon as she is found in heat, unless one knows the exact time of onset, in which case it would be best to wait 10–12 hours.

THE MARE

The diestrous cycle of the mare is generally given as 20–21 days, which is a good average, but there is a large spread in the curve of cycle lengths. Estrus lasts most often 3–5 days, but is sometimes shorter, more often longer (see table pp. 8 and 9 of Andrews and McKenzie, 1941).[19] Estrus is best determined by the use of a "teaser" stallion.

Ovulation in the mare has been determined with precision by rectal palpation of the ovaries (Mirskaia 1935;[60] Hammond, 1938b).[61] This procedure was, of course, for purposes of research only, as it is not a practical measure for the breeder. The time of ovulation may also be inferred on the basis of timed matings (Hammond, 1938a).[62] The Soviet workers, as well as Hammond of England, and Andrews and McKenzie are agreed that ovulation occurs 24–48 hours before the end of estrus.

There is little consolation in this conclusion for the horse breeder, however, for the preovulatory period is very variable (due, according to Hammond, to the variability of the different rates of growth of follicle in different mares). It may be stated, on the basis of known facts, that it is better not to breed the mare at the beginning of the heat period.

[59] Anderson, James. Emp. J. Exper. Agric. 4: 186–195. 1936.
[60] Mirskaia, L. M. Sechenov J. Physiol. U. S. S. R. 21: 195–196. 1935.
[61] Nalbandov, A., & F. C. Casida. J. An. Sci. 1: 189–198. 1942.
[62] Hammond, J. Yorkshire Agric. Soc. J. 95: 11–25. 1938(a).

Mirskaia and Petropavlovski (1937)[13] circumvented the dilemma by precipitating ovulation by means of human chorionic hormone. One thousand mouse units injected on the third or fourth day of heat may be expected to cause ovulation within 24 or 36 hours. Artificial insemination or mating would be indicated late on the same day, or better, on the day following the injection.

THE RHESUS MONKEY

From among the macaques indigenous to southeastern Asia, the rhesus monkey, *Macaca mulatta,* is the one which has established itself among us as a laboratory animal. I doubt whether the popularity of this species is due to superior qualities over others. It is probably its availability that is responsible. The animal is, however, hardy, long-lived and fertile in captivity.

For the taxonomic position and the wide geographical distribution of the races of the species, the reader is referred to the volume, "Anatomy of the Rhesus Monkey," Hartman and Straus, Ed., Williams and Wilkins, Baltimore, 1933. In the appendix of this volume will be found a description of various monkey quarters, including that of the Carnegie Laboratory of Embryology in Baltimore. Methods of handling monkeys, when kept in large paddocks, are also discussed.

As large males are dangerous to handle, it has been my practice to catch them only when they need to be transferred from one cage to another, treated for an injury, or the like. Instead, for breeding purposes, it is the females that are brought to the males. Appropriate sliding doors are provided for separating the two animals in order to remove the female.

Preparatory to mating, it is well to let the female become accustomed both to the cage and to the male or males to which she is to be bred; for, once she is frightened, she is likely to be attacked by the male and injured, which renders mating subsequently even more precarious. This difficulty is greatly lessened when only a few monkeys are confined in a room where they can see and "talk" to each other. Under such circumstances, Dr. Gertrude van Wagenen, at Yale University, has been very successful in breeding monkeys and rearing the young to maturity.

As I first found about 1930, it is possible to palpate the ovaries of the monkey and thereby determine with accuracy the time of ovulation (Hartman, 1930[62a]). Many hundreds of observations have been made, from which I selected 337, concerning which there was no element of

[62a] Hartman, Carl G. Anat. Rec. 45: 263. . 1930.

doubt as to the time of ovulation (Hartman, 1932,[9] 1936,[3] 1944[63]). In these, ovulation occurred 149 times on day 12 or 13, and 273 times between day 11 and day 14.

Around the middle of the cycle is, therefore, the optimum time to mate the monkey, if a pregnancy is desired.

The chimpanzee has a 36-day menstrual cycle, as compared with one of 28 days characteristic for the macaque female and for women (Hartman, 1936,[3] 1944[63]); and ovulates correspondingly later, i.e., around days 17–20 (See Yerkes, Chimpanzees, Yale University Press 1943, for references).

DISCUSSION OF THE PAPER

Dr. Robert K. Enders (*Swarthmore College, Swarthmore, Pa.*):

How fitting it is that this paper should be read by a man who has had such wide experience in mating mammals maintained in laboratory colonies, and that it should be read in a museum which was among the first to realize the need and develop the facilities for such colonies.

The wide taxonomic range that has been covered emphasizes the richness of the fauna that can be used in the laboratory. The work of Dr. Hartman has benefited by his knowledge of which species of mammals to use for a given research. It is only fair to assume that even greater riches will be uncovered as we become more familiar with a greater number of species. Thus, to mention but two new additions to the laboratory colony, the hamster and the cotton rat have opened fresh approaches to old fields of investigation. The discovery of still other mammals that can live under laboratory conditions is only a question of imagination and trial.

Dr. Hartman has spoken of the techniques of the animal breeder. The importance of technique is often overlooked and may result in failure or in data of doubtful value. In a statistical study of the breeding of mink, Apelgren (1941)[1] came to the conclusion that, as far as productivity was concerned, not more than 10 per cent was due to heredity, the remaining 90 per cent being due to external causes. Work by O. P. Pearson on the mink and fox indicates that at least 50 per cent of the variation in litter size is due to factors not genetic. These figures are based on large numbers of animals kept on different ranches and are mentioned here because they emphasize the need for just such information as has been given on mating, and they also emphasize the importance of technique in caring for animals.

While many breeders believe that the tendency to polygamous breeding by male foxes is inherited, I rather incline to the belief that this is training. If a male fox is permitted to spend the winter and the early part of the breeding season with one female, it may be difficult to induce him to breed with a strange vixen. On the other hand, if the dog fox has not been so treated, skillful handling will lead to willingness to mate with almost any estrous vixen.

There are few exceptions to the generalization that the use of hormones has not led to any increase in the overall fecundity of a herd or colony. The use of hormones in securing more pregnancies in the horse has been mentioned, and to this should be added the findings of Dryerre,[2] who reported that the production of a herd of foxes was increased by injecting prolan and thus bringing about, not greater production in the individual vixen, but an increase in the percentage of yearling vixens bred.

[63] Hartman, Carl G. West J. Surg., Obstet. & Gynec. (July) 52: 41–61. 1944.
[1] Apelgren, Hans. Kullstorleken och dräktighetstidens längd hos mink. Våra Pälsdjur, 12: 349–351, 6 figs. 1941.
[2] Dryerre, H. Prolan and fertility in silver fox. J. Physiol. 96: 35P–36P. 1939.

Dr. Frank A. Beach (*American Museum of Natural History, New York, N. Y.*):

Dr. Hartman quite properly stresses the specificity of the behavioral effects of ovarian and testicular hormones; but, if our thinking is to remain clear, we must be careful to avoid overemphasis upon this point. When administration of androgen induces masculine behavior in a genetic female, or when estrogen causes feminine responses in a male, how are we to interpret the phenomena? Dr. Hartman suggests that the heterologous hormone overrides the chromosomal contitution, and this view is entirely sufficient for practical purposes, but I prefer to think of the situation in somewhat different terms. Experimental data now available strongly suggest that the normal chromosomal constitution of the genetic male includes provision for the mediation of feminine sexual behavior; and a similar bisexual arrangement appears to exist in the genetic female. Administration of the heterologous hormone merely increases the reactivity of those mechanisms responsible for the overt responses typical of the opposite sex. The organization of the homologous mechanism is not affected, although its responsiveness to stimulation may be greatly reduced.

Along this same line it may be noted that the homologous or, perhaps we had better say, the *prepotent* behavioral mechanism may, under certain conditions, be activated by the heterosexual hormone. Thus, it is not strictly accurate to say that "estrogens alone are able to call forth mating behavior in the castrate female"; for the spayed female rat may be induced to receive the male, if she has been injected with androgen. Furthermore, the castrated male rat exhibits increased masculine sexual performance when injected with estrogen. It is of interest in this connection to note the reports of some clinicians to the effect that libido in the human female may be increased by testosterone propionate treatment more consistently than by ovarian hormone administration.

In speaking of the difficulties of increasing fertility by induced ovulation, Dr. Hartman mentioned the high incidence of "silent heat"—ovulation which is not accompanied by willingness to receive the male—and he observed that, "there is nothing that the animal breeder can do about it." I should like to add that the experimentalist interested in behavior should be able to do a great deal about it, and thus make a valuable contribution. Some headway in this direction has already been made in studies revealing different thresholds for the morphological and behavioral aspects of estrus. There is no reason why similar experimentation on a larger scale should not provide us with means of simultaneously inducing sexual receptivity and ovulation in domesticated species.

Dr. Hartman observed that male animals may be conditioned to mate with dummies in preference to a receptive female; and I am inclined to feel that this point deserves special attention because of its implications for practical animal breeding. It is well known that potentially valuable sires may be ruined for stud by improper handling during early services; and, conversely, the readiness with which a male attempts to mate with a female may be increased by appropriate conditioning in early adulthood. It has been repeatedly demonstrated that a male's sexual responsiveness can be either raised or lowered by previous experience, and regular schedules for the sexual conditioning of animals to be used at stud seem worthy of consideration.

One final point that may be mentioned rests upon the common observation that fertility, potency and libido are three separate functions which behave somewhat as independent variables. Thus a male or female may be fertile and yet fail to reproduce because of absence of sexual drive. Again, it is possible that a fertile male may never impregnate a female because of his impotence. Proper hormone treatment makes it possible, not only to increase fertility in the sense of inducing ovulation or spermatogenesis, but also to intensify the sex drive and thus induce or increase overt mating responses. Sexually sluggish but completely fertile males often mate readily and effectively following the administration of small doses of androgen—doses which are not damaging to the seminiferous tubules. Similarly, females showing socalled silent heat may be induced to receive the male and to conceive, if appropriate hormone administration is employed.

Dr. Edmond J. Farris (*The Wistar Institute, Philadelphia, Pa.*):

Dr. Hartman mentioned in the course of his paper that I should describe some of our findings on the study of reproduction in the albino rat. In our studies on reproduction in the female albino rat, a method was developed to establish, by running activity, puberty, estrus cycle regularity, four stages of estrus and menopause. With the aid of a blackboard, these findings can be quickly demonstrated. During estrus, in stage I, the female fought the male, and no intercourse occurred. In stage II of estrus (two hours after beginning of heat, as determined by the start of the females' running activity), 90 per cent of the resistant female rats became pregnant when permitted to mate no more than twenty minutes. The male ejaculated once at between the 15th to 18th minute of this twenty minute mating period. In stage III, both the male and the female mated willingly, and only 10 per cent pregnancies occurred. It required usually 55 to 80 minutes of mating before the male would ejaculate, during this stage. In stage IV, the male was uninterested in the female in heat.

It was established that ovulation occurred in Wistar albino rats eight hours after the beginning of heat in mated females, and slightly later in the non-mated.

Spermatozoa were identified in the uterus promptly after ejaculation. However, spermatozoa were seldom found in the oviduct or ovarian sac until just before or at time of ovulation.

Finally, the spermatozoa became immobile in the uterus four hours after ovulation, and disappeared completely from the uterus within three days after ovulation.

Dr. Arthur Zitrin and Dr. Frank A. Beach (*Department of Animal Behavior, American Museum of Natural History, New York, N. Y.*)[1]:

In the discussion of Dr. Hartman's paper on the mating of mammals, one of us mentioned the fact that we have been able to "condition" male cats so that they will mate frequently and consistently under laboratory conditions. The incidental statement called forth queries emphasizing the rather widely accepted belief that the domestic cat does not readily reproduce in the laboratory. This attitude is reflected in the reports of. Winiwarter and Saintmont (1908),[2] who state that cats will not breed under restraint.

Anestrous female cats show spontaneous courtship, mating and postcopulatory ovulation, following administration of appropriate gonadotropins (Friedgood, 1939[3]; Windle, 1939[4]); and spontaneous estrus occurs regularly in confined females. Therefore, any obstacle to the successful breeding of cats in the laboratory may be safely referred to some failure on the part of the male.

There is no reason to assume that close confinement affects the male's fertility. Although we have not studied in detail the gonads of males with which we worked, preliminary inspection of the sectioned testes from six males kept in the laboratory for more than a year indicates that spermatogenesis was normal. Ejaculates containing large numbers of motile sperm were obtained from one individual. Furthermore, the single intact female tested with our males became pregnant and delivered a normal litter. These observations, plus the fact that captive males invariably exhibited appreciable increase in body weight and gave other evidence of being in good health, lead us to suspect that the male cat's failure to impregnate the fertile females in the laboratory is due, primarily, to a lack of sexual aggressiveness. This impression is strengthened by the observations of Langley (1911),[5] who reports that "there is great variation in the behavior of males and that, while observers

[1] Observations here reported constitute part of an experimental investigation supported by a grant from the Committee for Research in Problems of Sex, National Research Council.

[2] **Winiwarter, H., & G. Saintmont.** Nouvelles recherches sur l'ovogenèse et l'organogenèse de l'ovaire des mammifères. Arch. de Biol. **24:** 1–47. 1908.

[3] **Friedgood, H. B.** Induction of estrous behavior in anestrous cats with the follicle-stimulating and luteinizing hormones of the anterior pituitary gland. Am. J. Physiol. **126:** 229–233. 1939.

[4] **Windle, W. F.** Induction of mating and ovulation in the cat with pregnancy urine and serum extracts. Endocrinol. **25:** 356–371. 1939.

[5] **Langley, W. H.** The maturation of the egg and ovulation in the domestic cat. Am. J. Anat. **12:** 139–172. 1911.

were present, few of those tried would act promptly, or at all. In a considerable number, only one was found which could be relied upon to afford a final test for oestrus."

In the course of an investigation into the effects of brain injury upon sexual activity, we have studied in some detail the mating performance of seventeen adult male cats. These animals were strays purchased from a source which supplies medical schools in the metropolitan area. Their history is unknown and the size and apparent age varied considerably. After a good deal of experimentation, a training schedule was devised which conditioned the cats so that 15 of the 17 cases mated promptly when confronted with an estrous female. Since all but one of the females used were spayed animals in which estrus was induced by estrogen treatment, pregnancy could not occur. However, as explained above, the males were almost certainly fertile and would have impregnated normal estrous queens.

In most cases, males showed no signs of mating until they had been in the laboratory for at least a month. Two individuals copulated (following training) after approximately two weeks' residence. A preliminary period of three to four weeks, in which the animals are allowed to become accustomed to feeding and being handled, was found to be an important prerequisite to successful sexual conditioning. After the acclimatization period, regular tests were conducted in which the following techniques proved helpful.

1. Only fully receptive females were used as stimulus animals. Tests for receptivity were brief and simple. When the loose skin on the back of the female's neck is grasped firmly in the experimenter's fingers, the receptive female will flex the forelimbs, lowering the anterior portion of the body. When the perineum is tapped gently, the receptive female exhibits treading responses, alternate stepping movements of the hind limbs. When a glass rod is inserted in the vagina the female frequently will growl loudly; and, after the neck grip is released and the glass rod withdrawn, the female displays the characteristic, violent after-reactions which include vigorous rolling and twisting on the floor, repeated licking of the vagina, etc.

2. Males were always taken from their living cages to a special test room, and they were never introduced into this room except upon the occasion of a sex test. Through this procedure, the males apparently came to associate the test room with sexual opportunity.

3. In beginning tests, males were allowed 15 minutes or longer to explore the test room before the stimulus female was introduced. In later tests, after conditioning had become apparent, this preliminary interval was found to be unnecessary.

4. The receptive female was placed in the test room and the observer took his station on the outside of the room where his presence did not distract the animals although their activity was clearly visible to him. In early tests, males responded to the female by thorough investigation, but there was little indication of sexual arousal, and mating was rarely attempted. As soon as it became apparent that the male was not going to copulate spontaneously, the experimenter reentered the test room and proceeded to stimulate the female to exhibit heat behavior.

5. Employing the techniques described above, the experimenter induced the female to crouch, tread, and display after reactions. After this performance had been repeated several times in succession, the male's interest in the female usually increased appreciably, although it was often necessary to continue stimulating the female for half an hour or longer before the male showed any evidence of arousal. Three males mated in their first test after the female had been induced several times to go through her entire mating role. The remaining 12 males did not copulate until this stimulus situation had been presented in several successive tests.

6. Conditioned males were tested regularly at least once each week. It is possible that less frequent testing might have been equally effective; but we found that weekly tests produced highly reliable results. Without exception, after a male mated in two or three preliminary tests, sexual responsiveness continued.

After they had copulated in several tests, the males became extremely reliable in their mating performance. They approached and mounted the female with very little delay; and the frequency of completed matings in a series of one hour tests was highly consistent for each individual.

It seems very likely that the simple procedures outlined above, or some variation of them, are customarily employed by commercial animal breeders; but, inasmuch as no published description of any such methods has come to our attention, it seems worth while to make our experience available to others who may be interested in observing behavior or inducing reproductivity in this species.

FEEDING LABORATORY ANIMALS

BY J. K. LOOSLI

Cornell University, Ithaca, N. Y.

Proper feeding of laboratory animals is an essential to maintaining a vigorous breeding stock useful in experimental investigation. In the efficient feeding of any species, it is helpful to understand the nutritional requirements for the various body functions, in order that foods may be selected to fully satisfy the varying needs.

More studies have been made with the rat than with any other laboratory animal and its nutritive requirements are more fully understood. Therefore, most space is devoted in this discussion to studies with rats. Limited attention is also given to the rabbit, guinea pig, mouse, hamster, and cotton rat. The nutrients required to prevent specific deficiency diseases are briefly considered and diets suggested which will maintain these animals with essential freedom from deficiencies during growth and reproduction.

NUTRITIONAL REQUIREMENTS

PROTEIN: Proteins are required for the formation of new tissues during growth, for replacement of protein tissues broken down in the normal metabolic processes, and for the synthesis of secretory products, such as milk. The amount of protein needed by an animal clearly varies with the type and extent of the metabolic function involved.

Hogan and Pilcher[1] observed more rapid gains in weight on high protein intakes than on a protein-low diet. Johnson *et al.*[2] also obtained more rapid gains in weight on a diet with 25 per cent protein than on one containing 10 per cent protein. The energy utilization was equally efficient at the two levels of protein intake, but on low-protein the animals stored more fat, while those on high-protein stored more protein and water. Forbes *et al.*[3] present data showing that a diet containing 20 per cent of protein produced more rapid gains in body weight and more efficient energy gain than lower levels, but 25 per cent of protein was not superior to the 20 per cent level. Hamilton[4] found that the growth rate of rats increased as the dietary protein was increased from 4 to 16 per cent, remained constant from 16 to 30 per cent, and decreased at higher levels of protein intake. The appetite of rats

[1] Mo. Res. Bull. **195.** 1933.
[2] Mo. Res. Bull. **246.** 1936.
[3] J. Nutr. **10:** 461. 1935.
[4] J. Nutr. **17:** 565, 582. 1939.

fed diets containing less than 16 or more than 22 per cent was adversely affected. He concluded that, for rats under 125 grams in body weight, the minimum protein requirement may be as high as 20 per cent of the dry matter.

McCoy[5] has presented a preliminary report of a study of the effect of different levels of protein in the diet upon growth, reproduction, lactation and body composition. On *ad libitum* feeding, young rats gained weight more rapidly when fed purified diets containing 25 to 40 per cent of protein than with 15 per cent protein. In paired feeding studies, the rate of growth and the percentage of nitrogen in the animal body paralleled the protein intake. In reproduction studies extending five generations, females on the high-protein diet produced more litters and more young per litter than those on the other diets As judged by the weight of the young at weaning, the diet containing 25 per cent protein produced the most satisfactory lactation

For normal growth of rats, the minimum protein appears to be supplied by a diet containing 20 per cent of protein to 125 grams in body weight, and 16 to 18 per cent thereafter. Levels of dietary protein of 30 per cent or higher may be disadvantageous for growth More data are needed on the requirements for lactation.

Amino Acids: Inasmuch as all tissue protein is made up of at least 22 generally recognized amino acids, it becomes obvious that all of these must be furnished either by synthesis within the body or supplied in the diet of the animal. Thus, the protein requirement actually becomes a need for specific amino acids, as illustrated by the early work of Osborne and Mendel.[6] They clearly demonstrated that the amino acids, lysine and tryptophane, are essential for the growth of rats.

Much of our knowledge regarding the amino acid requirements of rats has resulted from the studies of Rose and associates at Illinois. The results of these experiments demonstrate that 10 amino acids are indispensable for growth of the rat,[7] as shown in TABLE 1.

Fat: In 1929, Burr and Burr[8] first reported a deficiency disease produced by the exclusion of fat from the diet. They showed that, on a fat-low diet, rats develop a scaly condition of the skin, followed by a necrosis of the tail, a degeneration of the kidneys and finally a failure of growth and death. On a fat-low diet, females show irregular ovulation and males may refuse to breed. Hematuria, albuminuria and kid-

[5] J. Biol. Chem. 133: XIV. 1940.
[6] J. Biol. Chem. 17: 325. 1914.
[7] Physiol. Rev. 18: 109–136. 1938.
[8] J. Biol. Chem. 82: 345. 1929, 86: 587. 1930

TABLE 1
CLASSIFICATION OF THE AMINO ACIDS FOR GROWTH OF RATS

Essential	*Non-essential*
Lysine	Glycine
Tryptophane	Alanine
Histidine	Serine
Phenylalanine	Norleucine
Leucine	Aspartic acid
Isoleucine	Glutamic acid
Threonine	Hydroxyglutamic acid
Methionine	Proline
Valine	Hydroxyproline
Argenine*	Citrulline
	Tyrosine
	Cystine

* Argenine cannot be synthesized at a sufficiently rapid rate to permit normal growth.

ney injury are generally found. Burr and associates[9] found that linoleic and linolenic acids both cured the deficiency symptoms and, more recently, arachidonic acid has been shown to be effective (Turpeinen[10]). Evans and coworkers[11] found successful gestation was not possible on a fat-free diet. In pregnant animals, resorption may occur or the gestation period may be prolonged. Maternal mortality is frequent and any litters born are dead or so weak they soon die. Addition to the diet of large amounts of saturated fatty acids does not improve gestation, whereas unsaturated acids allow normal reproduction. Males develop a sterility which is cured or can be prevented by 50 mg. of unsaturated fatty acids prepared from corn oil.

Martin[12] has tentatively suggested that 30 mg. of methyl linoleate daily per rat will permit optimum growth. Turpeinen[13] obtained maximum growth response in plateaued fat-deficient female rats by feeding about 100 mg. methyl linoleate daily, or 33 mg. of the methyl ester of arachidonic acid. Nunn and Smedley-Maclean[14] observed that arachidonic acid is absent from the liver of rats fed a fat-deficient diet for several months, but it is present following the feeding of methyl linoleate. On a diet lower in fat than previously used, Mackenzie and associates[15] reported that, when 25 mg. of methyl linoleate was fed for as long as 11 months, the development of rats was comparable to those on a stock diet. Reproduction appeared normal, but many of the young died within 3 days after birth. The results of exchanging

[9] J. Biol. Chem. **97**: 1. 1932.
[10] J. Nutr. **15**: 351. 1938.
[11] J. Biol. Chem. **106**: 431, 441, 445. 1934.
[12] J. Nutr. **17**: 127. 1939.
[13] J. Nutr. **15**: 351. 1938.
[14] Biochem. J. **32**: 2178. 1938.
[15] Biochem. J. **33**: 935. 1939.

young from mothers fed low-fat and normal diets suggested that lactation on the low-fat diet is inadequate.

That fat in the diet improves lactation performance has been shown by Maynard and Rasmussen.[16] The beneficial effect of extra fat (corn or cottonseed oil) is evident, even when 125 mg. of ethyl linoleate are fed to the lactating mothers daily (Loosli and associates[17]), but no improvement in lactation performance was evident when fully hydrogenated coconut fat was fed.

Forbes and Swift[18] have shown that a certain amount of fat in the diets of rats is essential for a minimum energy loss as dynamic effect. The addition of fat to a protein-carbohydrate diet decreases the dynamic effect and thus increases the useful energy available to the animal. In the case of chickens (Russell *et al.*[19]), the addition of 4 per cent fat to a low-fat basal diet increased the utilization of carotene and permitted larger liver storage when vitamin A was fed. In rats, there is some evidence that the presence of fat in the diet facilitates utilization of carotene and also of calcium, but other studies have failed to confirm such an effect. The matter cannot be considered as settled.

Vitamin A and Carotene: Vitamin A is essential for all mammals. It is important for growth, normal vision and the maintenance of normal epithelial tissues. Vitamin A deficiency results in night blindness, keratinization of epithelia, a disturbance of bone growth, sterility and cessation of growth.

Wolback and Howe[20] first clearly demonstrated that lack of vitamin A causes replacement of many different epithelia by stratified keratinized epithelium. Functional failure of glands may result. In the female, sterility may ensue from implantation failure due to keratinization of the uterine epithelium (Evans[21]). Fridericia and Holm,[22] Tansley[23] and others have reported that night blindness is one of the early symptoms of vitamin A deficiency. Mason[24] has shown that vitamin A is necessary for maintenance of normal epithelium in the testis. Mellanby[25] has presented evidence indicating that vitamin A deficiency in young dogs causes overgrowth and deformities of the bones. The overgrowth of bones causes pressure upon nervous tissue indirectly producing degenerative changes which were earlier thought

[16] J. Nutr. 23: 385. 1942.
[17] J. Nutr. 28: 81. 1944.
[18] J. Nutr. 27: 453. 1944.
[19] J. Nutr. 24: 199. 1942.
[20] J. Exp. Med. 42: 753. 1925.
[21] J. Biol. Chem. 77: 651. 1928.
[22] Am. J. Physiol. 73: 63, 79. 1925.
[23] J. Physiol. 71: 442. 1931.
[24] Am. J. Anat. 52: 153. 1933.
[25] J. Physiol. 94: 380. 1938.

to be a direct effect of vitamin A lack (Irving and Richards[26]). Wolbach and Bessey[27] have shown that in vitamin A deficiency in rats, skeletal growth is retarded earlier than that of the soft tissues in general, including that of the central nervous system, and that the nervous manifestations are due to pressure effects caused by relative overgrowth of the central nervous system. Evidence that a stenosis of the optic canal causes constriction of the optic nerve and blindness in vitamin deficiency of young calves is presented by Moore.[28] Wolbach and Howe[29] and Orten *et al.*[30] show that vitamin A deficiency produces upon the incisor teeth a loss of normal orange pigment, opacity, distortion of shape and exfoliation. Mellanby[31] observed pale and deformed teeth in young rats reared by mothers kept for several months on a diet low in vitamin A. A detailed discussion is given by Schour and Massler.[32]

Vitamin A plays a part in the metabolism of visual purple (Wald[33]). Johnson[34] has shown that severe vitamin A deficiency, in addition to producing night blindness, may result in a structural breakdown of the retina itself in rats.

In 1928, Euler[35] reported that 5 micrograms of carotene daily was enough to permit a resumption of growth in rats suffering from vitamin A deficiency. Moore[36] showed that, as a source of vitamin A, carotene was effective at a level of 4 micrograms daily. Guilbert *et al.*[37] have shown that the minimum requirements of the rat are 18 to 22 units of vitamin A, or 15 to 20 micrograms of carotene per kilogram of body weight. This is about in line with several other species of animals studied, suggesting that vitamin A is concerned with general metabolic activities of the body as a whole. Goss and Guilbert[38] observed that 18 to 22 units (3.8 to 4.6 micrograms) of vitamin A, or 15 to 20 micrograms of carotene per kilo of body weight daily, is necessary to prevent cornification of the vaginal epithelium. No storage of vitamin A in the liver occurred until 80 micrograms of carotene per kilo of body weight were given daily. Cannon[39] has reported that female rats de-

[26] J Physiol. **94:** 307. 1938.
[27] Am. J. Path. **17:** 586. 1941.
[28] J Nutr. **17:** 443. 1939.
[29] Am. J. Path. **9:** 275. 1933.
[30] Proc. Soc. Exp. Biol. Med. **36:** 82. 1937.
[31] Brit. Dental J. **67:** 187. 1939.
[32] The Rat in Laboratory Investigation. Griffith and Farris, Eds., J. B. Lippincott Co., Philadelphia. 1943.
[33] Proc. Nat. Acad. Sci. **25:** 344. 1939.
[34] J. Exp. Zool. **81:** 67. 1939.
[35] Biochem. Z. **203:** 370. 1928.
[36] Biochem. J. **24:** 692. 1930.
[37] J. Nutr. **19:** 91. 1940.
[38] J. Nutr. **18:** 169. 1939.
[39] Proc. Soc. Exp. Biol. Med. **44:** 129. 1940.

pleted of vitamin A until they showed xerophthalmia or weight loss generally did not mate. Others on the depletion diet, plus 40 micrograms of carotene three times weekly, mated and delivered living young. Lewis *et al.*[40] found 50 I.U. daily were necessary to give a significant amount of vitamin A in the liver of rats. None was found on daily intakes of 10 I.U. or less.

VITAMIN D: Vitamin D is related to the utilization of calcium and phosphorus, and it is essential for normal calcification of the growing bone. No amount will compensate for severe deficiencies of calcium or phosphorus, but it helps to alleviate the effect of a wide Ca:P ratio.

Schneider and Steenbock[41] observed that, on a low phosphorus diet (0.57 per cent calcium, 0.04 per cent phosphorus) free from vitamin D, young rats gained weight for 4 to 5 weeks, developed severe rickets and bone deformities and died after 6 weeks. Adding vitamin D caused weight loss when the young were placed upon the diet, and gains were very slow. From tissue analyses and calcium and phosphorus balances, the authors suggest that vitamin D induces utilization of phosphorus for bone growth, thereby depriving the soft tissues of a phosphorus supply, which in turn inhibits growth. Templin and Steenbock[42] have furnished evidence that vitamin D markedly decreases the losses of ash from the bones that otherwise occur in the adult rat on a diet very low in calcium. Nicolaysen[43] has suggested that, in vitamin D deficiency, a decreased calcium absorption results. The increased fecal calcium results in precipitation of phosphorus, rendering it unavailable. Cohn and Greenberg[44] have reported that the absorption of phosphate by rachitic rats is increased only slightly by vitamin D, but the phosphorus uptake in the bone is increased by 25 to 50 per cent. Cox and Emboden[45] have shown with rats that growth, bone calcification and reproduction are excellent when adequate amounts and favorable ratios of calcium and phosphorus are available, despite a very low intake or absence of vitamin D in the diet.

VITAMIN E: Vitamin E is known to be essential for normal reproduction in the rat, mouse, and in poultry. When the factor is absent from the diet of the female, death and resorption of the embryos result. Ovulation and conception is not interfered with. In the male rat, permanent sterility results from degenerative changes of the epithelium

---

[40] J. Nutr. 23: 351. 1942.
[41] J. Biol. Chem. 128: 159. 1930.
[42] J. Biol. Chem. 100: 209. 1933.
[43] Biochem. J. 31: 105, 107, 122. 1937.
[44] J. Biol. Chem. 130: 625. 1939.
[45] J. Nutr. 11: 147. 1936.

of the testes. The early studies have been reviewed by Evans.[46] Mason[24] has carefully differentiated the pathology in vitamin E deficiency from that produced by a lack of vitamin A. Many workers have shown that the growth rate of rats is retarded on a vitamin E deficient diet.[47]

Rabbits and guinea pigs fed diets containing cod liver oil develop a muscular dystrophy (Madsen *et al.*[48]) which can be cured or prevented by feeding a-tocopheral (Eppstein and Morgalis[49]). A similar dystrophy is shown by rats on vitamin E-low diets, particularly the young suckled by vitamin E-low mothers. Mason and Bryan[50] have shown that the placental transfer of vitamin E is very limited, but that larger amounts pass into the milk.

Evans and Emerson[51] have shown that the amount of vitamin E required by rats for normal reproduction increases with age. Feeding 0.10 mg. of alpha-tocopheral acetate, six times weekly, maintained normality of striated musculature in both males and females. In males, 0.10 mg. would not maintain fertility more than five months, and 0.25 mg. was inadequate after nine months, while 0.75 mg. maintained normal testes and fertility during the sixteen month experiment. Females on 0.10 mg. produced living young for three gestations (11 to 12 months). By eighteen months, orange-brown pigmentation of the uterus occurred in females receiving 0.10 and 0.25 mg. but not in those given 0.75 mg. of alpha-tocopheral acetate six times weekly. Young born of mothers on the 0.10 mg. level exhibited muscle dystrophy and high mortality. On 0.25 mg., the incidence was less until the third gestation, when all young were paralyzed. On 0.75 mg., all young were normal for two litters, but, in the third gestation, a slight temporary dystrophy was seen.

Eppstein and Morgulis[49] have reported that a daily dose of 0.32 mg. of alpha-tocopheral per kilo of body weight will cure muscle distrophy in rabbits.

In the mouse, the administration of 0.5 to 1.0 mg. alpha-tocopheral at the beginning of gestation resulted in normal young in at least 85 per cent of the cases (Goettsch[52]), whereas those without the supplement produced no litters. The second generation males on the vitamin

[46] J. Am. Med. Assoc. **99**: 469. 1932.
[47] Cited, Ref. [33].
[48] Cornell Univ. Agr. Exp. Sta. Memoir: **178**. 1935.
[49] J. Nutr. **22**: 415. 1941.
[50] J. Nutr. **10**: 501. 1940.
[51] J. Nutr. **26**: 555. 1943.
[52] J. Nutr. **23**: 513. 1942.

tion of a red, fluorescent, porphyrin-containing exudate on the nose, whiskers and fur.

Schaefer *et al.*[76] have produced acute pantothenic acid deficiency in dogs, that is characterized by sudden collapse associated with decreased blood dextrose, increased nonprotein nitrogen and lowered blood chlorides. Severe intussusception in the intestinal tract and fatty livers have also been observed. Phillips and Engle[77] earlier reported specific neuropathologic changes in the spinal cord in chicks suffering from pantothenic acid deficiency. Wintrode *et al.*[78] have found nerve degeneration in pigs fed synthetic diets low in pantothenic acid and other members of the B complex.

The pantothenic acid requirement of dogs is about 0.10 mg. per 100 grams ration (Elvehjem[79]). Sixty micrograms of calcium pantothenate per day prevented graying in rats, but 100 micrograms were necessary for maximum growth (Henderson *et al.*[80]).

Leonards and Free[81] have observed that the rate of intestinal absorption of galactose was 15 per cent more in rats on normal diets than in animals deficient in pantothenic acid. Supplementation of a normal diet with 300 and 1000 micrograms of calcium pantothenate had no effect on the rate of absorption.

The biochemistry of pantothenic acid has been reviewed by Williams.[82]

PYRIDOXINE: Pyridoxine has been shown to be essential in the diet of the rat,[83] the dog,[84] the pig and poultry. A lack of this vitamin causes retarded growth, an acrodynia characterized by edema, swelling and denuding of the paws and areas around the mouth, and frequently thickening of the ears. A microcytic hypochromic anemia is characteristic in dogs. Sullivan and Evans[85] have described the dermal changes as distinguished from those seen in magnesium deficiency. Patton *et al.*[86] have reported the occurrence of convulsive seizures in young rats suckling mothers fed diets deficient in pyridoxine. The literature dealing with dermal changes in deficiencies of pyridoxine and the essential fatty acids is reviewed by McCoy.[32]

[76] J. Biol. Chem 143: 321. 1942.
[77] J. Nutr. 18: 227. 1939.
[78] Bull. Johns Hopkins Hosp. 67: 377. 1940.
[79] Hand Book of Nutrition. Am. Med. Assoc., Chicago. 1943.
[80] J. Nutr. [illegible]: 47. 1942.
[81] J. Nutr. [illegible]: 403. 1942.
[82] Advances in Enzymol. 3: 253. 1943.
[83] György et al. Proc. Soc. Exp. Biol. Med 37: 313. 1937.
[84] J. Nutr. [illegible]: 197. 1938; J. Nutr. 21: 275. 1941; J. Biol. Chem. 142: 77. 1942.
[85] J. Nutr. 27: 123. 1944.
[86] J. Biol. Chem. [illegible]: 121. 1943.

Emerson and Evans[87] have reported uniformly defective sexual behavior in pyridoxine deficient rats. Sure[72] found 10 to 25 micrograms of pyridoxine adequate for growth, but at least 50 micrograms for satisfactory lactation in rats.

BIOTIN: When young rats are fed a diet containing uncooked egg white as the source of protein, there develops a disease characterized by poor growth, generalized pruritic, exfoliative dermatitis, abnormal posture, abnormal gait and hypertonicity. The cutaneous signs described by Boas[88] and later workers are "spectacle eye," progressing to general alopecia, onset of a spasticity and finally death. Lease *et al.*[89] found that the rabbit and the monkey also exhibited a characteristic dermatitis when fed rations rich in egg white. Du Vigneaud *et al.*[90] first identified biotin as the vitamin involved, and Gyorgy *et al.*[91] established the presence of "avidin" as the biotin inactivating factor in egg white. Sullivan *et al.*[92] have suggested that the hypertonicity may be due to irritability caused by severe pruritus, since no anatomical changes in the central nervous system were evident.

Nielson and Elvehjem[93] found that, on an egg white diet, 2 micrograms of biotin per rat daily prevented or cured the "spectacle eye" syndrome. Signs of biotin deficiency do not appear on synthetic diets without egg white.

NIACIN (NICOTINIC ACID): The dog, pig and monkey are the only experimental animals that show typical nicotinic acid deficiency. Niacin is not required, preformed in the diets of rats or chicks (Harris and Raymond[94]), Axelrod *et al.*,[95] Sure.[96] Elvehjem *et al.*[97] showed that blacktongue in dogs could be prevented or cured by niacin. That the silver fox also requires niacin has recently been shown (Hodson and Loosli.)[98]

Nicotinic acid functions as a component of two important coenzymes —coenzyme I, or cozymase, and coenzyme II—which are concerned in both glycolysis and respiration (Elvehjem[99]). In nicotinic acid deficiency, there is a decreased cozymase content of the liver and muscle tissue.

[87] Am. J. Physiol **120**: 352. 1940.
[88] Biochem. J. **18**: 422. 1924.
[89] Biochem. J. **31**: 488. 1937.
[90] Science **92**: 62. 1940.
[91] Science **93**: 477. 1941.
[92] Bull. Johns Hopkins Hosp. **70**: 177. 1942.
[93] Proc. Soc. Exp. Biol. Med. **48**: 349. 1941.
[94] Biochem. J. **33**: 2037. 1939.
[95] J. Biol. Chem. **131**: 85. 1939.
[96] J. Nutr. **19**: 57. 1940.
[97] J. Biol. Chem. **123**: 137. 1938.
[98] Vet. Medicine **37**: 470. 1942.
[99] Physiol. Rev. **20**: 249. 1940.

VITAMIN K: A deficiency of vitamin K is manifested by a marked tendency to severe hemorrhage associated with an increase in the blood-clotting time. The symptoms are prevented or cured by vitamin K. Several investigators have shown the synthesis of vitamin K by the intestinal flora of rats (Dam *et al.*[100]; Greaves and others[101]), thus explaining the usual failure to produce the deficiency in this species. Greaves found that only 12 of 77 animals reared and maintained on a vitamin-K-free diet exhibited subnormal prothrombin values and hemorrhagic tendencies. Bile-fistula or jaundiced animals showed marked bleeding tendencies associated with low prothrombin, which was cured by feeding vitamin K. Day *et al.*[102] have presented data showing the cecum of the rat is an important site of vitamin K synthesis, but that this vitamin can also be formed in other parts of the intestinal tract. Cecectomized rats showed very low incidence of hypoprothrombinemia, whereas operated animals fed one per cent sulfasuxidine, exhibited a very high incidence. On the sulfasuxidine diet, the addition of p-aminobenzoic acid markedly reduced the incidence of hypoprothrombinemia, thus counteracting the effect of sulfasuxidine on vitamin K synthesis in the intestinal tract. This is in agreement with earlier studies (Elvehjem[103]).

CHOLINE: For many years choline has been known as a component part of the phospholipid lecithin, but its functional importance in nutrition was not apparent until Best and Huntsman[104] demonstrated its role in the prevention of fatty livers.

Choline is now generally considered an important member of the vitamin B complex, although Jacobi *et al.*[105] have demonstrated that it may be synthesized in the rat. The function of choline is related to the mobilization of fatty acids in the body. Du Vigneaud *et al.*[106] observed that the methyl groups of choline as well as those of methionine and betaine are transferable in the animal body. McHenry[107] states that choline may function in at least three ways: to stimulate the formation of phospholipids, to make possible the production of acetyl choline, or to supply labile methyl groups.

Griffith[108] reported fatty degeneration of the liver, hemorrhagic renal lesions, ocular hemorrhages and regression of the thymus within ten

[100] Biochem. J. 31: 22. 1937.
[101] Am. J. Physiol. 125: 428, 429. 1939.
[102] J. Nutr. 25: 585. 1943.
[103] Chem. and Eng. News 21: 853. 1943.
[104] J. Physiol. 75: 405. 1932.
[105] J. Biol. Chem. 138: 571. 1941.
[106] Biological Symposia 5: 334. 1941.
[107] Biological Symposia 5: 177. 1941.
[108] J. Nutr. 22: 239, 1941; 19: 487. 1940; J. Biol. Chem. 131: 567. 1939.

days after the rats had been placed on a choline-low but otherwise efficient diet. He found 1.0 mg. of choline chloride daily was needed to prevent renal lesions in young rats and 2 to 3 mg. to prevent fatty livers. Sure[109] reported 15 mg. choline chloride usually gave satisfactory lactation.

ASCORBIC ACID: Deficiencies of ascorbic acid can be demonstrated only in man, monkey and the guinea pig. Other species have the ability to synthesize this vitamin. The pathology of vitamin C deficiency has been reviewed in detail by Dalldorf.[110] Five to ten milligrams of ascorbic acid appear to be sufficient to protect the guinea pig against symptoms of scurvy.

OTHER VITAMINS: Several groups of workers have recently shown that the guinea pig requires two or more dietary factors in addition to all the known vitamins discussed in this paper and also inositol, folic acid and p-amino benzoic acid. Hogan and Hamilton[111] obtained fair growth and survival of guinea pigs fed purified diets supplemented with the known vitamins and factors from dried yeast and liver. Sober *et al.*[112] found that essential unknown vitamins are supplied by yeast, dried grass and whole milk. Wooley[113] and Kuiken *et al.*[114] show that guinea pigs need three factors in addition to the known vitamins. The rat apparently does not require a dietary source of these factors.

CALCIUM AND PHOSPHORUS: Calcium and phosphorus are discussed together because of the interrelationships of these mineral elements in nutrition. This relationship is well illustrated by the report of Bethke, Kick and Wilder[115] who suggest the Ca:P ratio is more important in determining the adequacy of the intakes than are the actual amounts of the elements. As they increased the Ca:P ratio from 1:1 to 5:1, a progressive decrease occurred in growth, bone ash and the inorganic phosphorus in the blood. Decreasing the Ca:P ratio from 1:1 to 0.25:1 decreases the growth rate and the serum calcium, but has only a slight effect upon the bone ash. They observed best growth when dietary calcium was 0.69 per cent and phosphorus 0.56 per cent. Cox and Imboden,[116] however, found that successful reproduction and lactation in the rat were dependent upon both the actual level and the Ca:P ratio in the diet. The best results were obtained on a diet containing 0.59 per cent of each calcium and phosphorus (a daily intake

[109] J. Nutr. 19: 71. 1940.
[110] The Vitamins. Am. Med. Assoc., Chicago. 1939.
[111] J. Nutr. 23: 533. 1942.
[112] J. Nutr. 24: 503. 1942.
[113] J. Biol. Chem. 143: 679. 1942.
[114] J. Nutr. 27: 385. 1944.
[115] J. Biol. Chem. 98: 389. 1932.
[116] J. Nutr. 11: 147. 1936.

of 42 mg.). Reproduction was poor when the diet contained 2.45 per cent of either element. Hubbell *et al.*[117] have suggested a salt mixture which allows rapid growth with an average daily intake of 50 mg. calcium and 35 mg. phosphorus. Sherman and Booher[118] and Sherman and Campbell[119] reported that, with diets containing 0.42 per cent phosphorus, increasing the calcium progressively from 0.16 to 0.50 per cent permitted more rapid calcification of bones, earlier maturity, and deferred senescence. Van Duyne *et al.*[120] found diets containing 0.64 to 0.80 per cent calcium optimum in promoting nutritional well-being as shown by full-life experiments in three generations.

Day and McCollum[121] reported that young rats on a diet containing 0.4 per cent calcium and 0.017 per cent phosphorus, but otherwise adequate, grow slowly for 5 to 6 weeks, then decline in weight and soon die. Extreme rarefraction of the bones occurs, accompanied by progressive disability in walking, standing and breathing. A loss of calcium occurs in the urine. A smaller loss of phosphorus (largely in the urine) suggests that part of this element mobilized from the bones is used by the soft tissues for growth. This severe phosphorus deficiency does not have any significant effect upon the metabolism of sodium, potassium, or magnesium. Boelter and Greenberg[122] found rats reared from weaning on a low-calcium diet (10 mg. per 100 gm. food) failed to mate, and mothers on calcium-low diet did not lactate normally.

Most satisfactory growth, reproduction and livability are obtained when the diet of rats contains about 0.6 per cent calcium and 0.5 per cent phosphorus, although good results can be obtained at appreciably lower levels.

MAGNESIUM: Sullivan and Evans[123] have recently produced uncomplicated magnesium deficiency in rats. Two or three days after weaning rats were placed on the magnesium-deficient diet, they were nervous and hyperirritable. After the first week, spontaneous or induced convulsions occurred. Some animals succumbed immediately after the convulsions, others survived several seizures. The first cutaneous sign was erythema. The paws soon became swollen and red. Growth was normal for the first two or three weeks. As the deficiency progressed, the animals became weak, less irritable and listless. After the sixth to eighth week, there was nutritive failure and variable superimposed

117 J. Nutr. 14: 273. 1937.
118 J. Biol. Chem. 93: 93. 1931.
119 J. Nutr. 10: 363. 1935.
120 J. Nutr. 21: 221. 1941.
121 J. Biol. Chem. 130: 269. 1939; J. Nutr. 20: 181. 1940.
122 J. Nutr. 26: 105. 1943.
123 J. Nutr. 27: 123. 1944.

illness. Occasionally, there was loss of hair and dermatitis. The dermal changes are distinguishable from pyridoxine deficiency. There is no relation between magnesium deficiency and that of the vitamin B complex.

These gross symptoms are generally similar to those earlier described by Kruse, Orent and McCollum[124] in rats and dogs; Tufts and Greenberg[125] and Watchorn and McCance[126] in rats.

The serum magnesium falls in magnesium deficient animals, as does the percentage of magnesium in the skeleton and soft tissues. The calcium content of the soft tissues and bones increases. Calcification of the kidneys and blood vessels have been reported. Snyder and Tweedy[127] have found that there is a marked fall in serum phosphatase activity associated with the decline in magnesium. There is a disturbance in the calcification of the teeth (Becks and Furuta[128] and Irving[129]).

Tufts and Greenberg[130] found that 5 mg. of magnesium per 100 grams of diet (4 mg. per kilo body weight) is about the minimum requirement for growth. Reproduction was normal at this level of intake, but suckling young showed signs of deficiency.

MANGANESE: In 1931, it was shown that manganese was essential in the diet of the rat (Orent and McCollum[131]) and the mouse (Kemmerer *et al.*[132]). Since that time, manganese has been reported to be essential for normal reproduction, lactation, bone formation, growth and the activity of certain enzymes. Workers are agreed that manganese deficiency produces testicular atrophy, sterility and retarded growth. Kemmerer *et al.*,[132] Waddell *et al.*,[133] Skinner *et al.*[134] and Boyer *et al.*[135] report disturbances of the estrous cycle resulting in sterility, whereas Orent and McCollum,[136] Daniels and Everson[137] and Shils and McCollum[138] observed no such effect. Manganese is important for normal bone development in chicks (Wilgus *et al.*,[139] and Caskey[140]). The

124 J. Biol. Chem. **96**: 519. 1932; Am. J. Physiol. **101**: 454. 1932.
125 J. Biol. Chem. **122**: 693. 1937–38.
126 Biochem. J. **31**: 1379. 1937.
127 J. Biol. Chem. **146**: 639. 1942.
128 J. Am. Dent. Assoc. **26**: 883. 1939.
129 J. Physiol. **99**: 8. 1940.
130 J. Biol. Chem. **122**: 715. 1937–38.
131 J. Biol. Chem. **92**: 651. 1931.
132 J. Biol. Chem. **92**: 623. 1931.
133 J. Nutr. **4**: 53. 1931.
134 Am. J. Physiol. **101**: 591. 1932.
135 J. Biol. Chem. **143**: 417. 1942.
136 J. Biol. Chem. **98**: 101. 1932.
137 J. Nutr. **9**: 191. 1935.
138 J. Nutr. **26**: 1. 1943.
139 J. Nutr. **14**: 155. 1937.
140 Proc. Soc. Exp. Biol. Med. **44**: 332. 1940.

studies of Barnes *et al.*[141] and Shils and McCollum[142] suggest the same is true for rats. Boyer *et al.*[135] reported a decreased arginase activity in manganese deficiency. Shils and McCollum[142] found manganese-deficient young may show loss of both equilibrium and coordination. High dietary calcium accentuates the severity of manganese deficiency in both chicks and rats. High intakes of manganese retards the growth rate of rats.[142] Adding 0.005 to 0.05 per cent of manganese to a low-manganese diet maintained fertility in males.[136] A daily intake of 0.5 to 0.8 mg. appears to be adequate for reproduction.

Smith *et al.*[143] have shown that manganese deficiency in rabbits results in a decreased breaking strength, weight, density, length and ash content of the humeri. Deformed front legs is the most evident gross symptom. The manganèse requirement for growth of rabbits appears to be about 1.0 mg. daily.

IRON AND COPPER: The history of the development of our knowledge regarding the role of iron in hemoglobin formation has been reviewed by Robscheit-Robbins.[144] Hart *et al.*[145] first reported evidence that copper was also essential for hemoglobin formation.

Josephs[146] reported that copper does not effect iron retention, but is concerned in hemoglobin formation. Elvehjem and Sherman[147] and Cunningham[148] found that, when iron is fed to the anemic rat in the absence of copper, the iron content of the liver and spleen increases but there is no rise in the hemoglobin level. Stein and Lewis[149] concluded that copper has two roles in blood formation: a catalytic action on hemoglobin formation and a stimulating effect on erythropoiesis. Cohen and Elvehjem[150] presented evidence suggesting that copper was concerned in cytochrome formation. They also showed that feeding copper or copper and iron markedly increased the oxidase content of the livers of anemic rats, but iron alone had no effect. Smith and Medlicott[151] have recently reported that deficiency of iron or copper, or both, produces a microcytic hypochromic anemia. Feeding pure iron to milk-anemic rats leads to a significant increase in the mean cell volume. Feeding copper to milk-anemic rats produced a rise in the erythrocyte count which was not accompanied by an increase in hemoglobin. This

141 Proc. Soc. Exp. Biol. Med. 46: 562. 1941.
142 J. Nutr. 23: 445. 1942.
143 Arch. Biochem. 4: 281. 1944; and unpublished data.
144 Physiol. Rev. 9: 666. 1929.
145 J. Biol. Chem. 77: 797. 1928.
146 J. Biol. Chem. 96: 559. 1932.
147 J. Biol. Chem. 98: 309. 1932.
148 Biochem. J. 25: 1267. 1931.
149 J. Nutr. 6: 465. 1933.
150 J. Biol. Chem. 107: 97. 1934.
151 Am. J. Physiol. 141: 354. 1944.

evidence corroborates the view of Stein and Lewis[149] that copper has a stimulating effect on erythropoiesis or on the release of erythrocytes from the bone marrow.

Mitchell and Miller[152,154] found 0.25 mg. of iron daily maintained normal hemoglobin when sufficient copper was present and that with less than 0.1 mg. of copper hemoglobin synthesis was retarded.

Hart *et al.*[145] have shown that iron and copper are the only constituents of liver needed for maximum hemoglobin formation by milk anemic rats.

There is evidence[152,153,154,155,156] that 0.25 mg. of iron and about 0.05 mg. of copper per day is adequate to maintain hemoglobin values or to regenerate hemoglobin rapidly in young anemic rats. A daily intake of 5 mg. of iron and 0.5 mg. of copper appears adequate for normal reproduction and lactation (Daniels and Everson[157]) but these values are probably more than the minimum requirements.

POTASSIUM: A detailed picture of uncomplicated potassium deficiency in rats has been presented by the studies of Orent-Keiles and McCollum.[160] When young rats were fed a diet containing only 0.01 per cent of potassium, growth continued, but at a slow rate, and the length of life did not seem to be affected. There was a roughness and thinning of the fur, and the animals showed an extreme alertness and marked pica. Ovulation was irregular and at a slower rate with occasional cessation of estrus. Sexual maturity was delayed. The potassium in the muscles, heart, and kidneys was lowered, while sodium increased. All of the potassium-deficient animals exhibited necrosis of the cardiac musculature and damage to the renal tubular epithelium, but other tissues were normal according to a report by Follis *et al.*[161] Cardiac and renal hypertrophy was observed. These studies confirm earlier experiments with rats by Schrader *et al.*[162] and by Thomas *et al.*[163] that cardiac and renal injuries result from potassium deficiency. According to Heppel,[164] serum potassium falls to about one half normal in potassium-deficient rats, and the chloride is about 15 per cent lower.

Heppel and Schmidt[165] found that pregnancy was possible when the diets of rats contained 0.010 to 0.015 per cent of potassium, but the

152 J. Biol. Chem. **92:** 421. 1931.
153 J. Biol. Chem. **104:** 217. 1934.
154 J. Biol. Chem. **85:** 855. 1929–30.
155 J. Nutr. **4:** 469. 1931.
156 J. Nutr. **5:** 295. 1932; **23:** 47. 1942.
157 J. Nutr. **9:** 191. 1935.
160 J. Biol. Chem. **140:** 337. 1941.
161 Am. J. Path. **18:** 29. 1942
162 J. Nutr. **14:** 85. 1937
163 Yale J. Biol. Med. **12:** 345. 1940.
164 Am. J. Physiol. **127:** 385. 1939.
165 Univ. of Calif. Pub. Physiol. **8:** 189. 1938.

mothers lost body tissue and the young were eaten at birth. During the lactation period, a dietary content of 0.58 per cent of potassium usually resulted in some storage. Miller[166] earlier stated the minimum potassium requirement for normal growth of rats was about 15 mg. for males and 8 mg. for females (0.55 to 1.44 gm. potassium per kilo of ration). The maintenance requirement is not more than 2 mg. daily.

SODIUM AND CHLORINE: Orent-Keiles *et al.*[167] found that retardation of growth, disturbances in reproductive functions, lesions of the eyes, and death resulted when rats were restricted to a diet which contained only 0.002 per cent of sodium. A detailed, histological study of Follis *et al.*[168] failed to reveal any specific tissue changes other than those in the ocular apparatus.

On a sodium-deficient diet, rats were in negative sodium balance, but retention of potassium and magnesium increased so that the acid-base balance was not appreciably disturbed (Orent-Keiles and McCollum[169]). Kahlenberg *et al.*[170] have shown that sodium deprivation did not affect the ability of rats to digest and absorb protein and energy. There was, however, a depressed appetite, increased heat production, and decreased body storage of protein and energy. A similar depression of the utilization of protein and energy has been shown to occur when rats are fed a diet deficient in chloride (Voris and Thacker[171]). Differing from sodium deficiency, however, the drastic restriction of chloride intake does not appear to cause any pathological tissue changes nor death. Marquis,[172] and Greenberg and Cuthbertson[173] reported that rats continued to grow at a reduced rate on a chloride-low diet. The latter authors showed that the urinary excretion of chloride fell rapidly and that rats receiving only about 1 mg. of chloride daily were still in positive chlorine balance.

COBALT: It has not been definitely established that cobalt is required by rats or other laboratory animals, although such a need may be indicated on the basis of studies with sheep and cattle. If rats have a need for cobalt, the amount required is less than 0.4 microgram daily (Underwood [174]). Orten *et al.*,[175] Stare and Elvehjem[176] and Meyer

[166] J. Biol. Chem. **55**: 61. 1923; **70**: 587. 1926.
[167] Am. J. Physiol. **119**: 651. 1937.
[168] Arch. Path. **33**: 504. 1942.
[169] J. Biol. Chem. **133**: 75. 1940.
[170] J. Nutr. **13**: 97. 1937.
[171] J. Nutr. **23**: 365. 1942.
[172] Compt. rend. soc. biol. **128**: 449. 1938.
[173] J. Biol. Chem. **145**: 179. 1942.
[174] Nutr. Abs. and Rev. **9**: 515. 1940.
[175] J. Biol. Chem. **96**: 11. 1932; **99**: 457. 1932–3; Am. J. Physiol. **114**: 414. 1935–6.
[176] J. Biol. Chem. **99**: 473. 1932–3.

et al.[177] have shown that feeding cobalt to rats (0.5 mg. daily) produces a marked polycythemia. It has been suggested that the cobalt produces an increase in the rate of formation of hemoglobin and erythrocytes.

ZINC: Studies at Wisconsin have shown that zinc is necessary for the growth of rats. On a nearly zinc-free diet, growth is retarded, and the development of the fur is interfered with (Stirn *et al.*[178]). The amount of zinc in the body, and especially in the bone, teeth and blood is reduced. Hove *et al.*[179] found that 40 micrograms of zinc per day prevented the deficiency.

IODINE: Iodine is essential in the diet in order to prevent enlargement of the thyroid gland (goiter). Remington and associates[180] have studied the iodine requirement of rats. The thyroid greatly increases in total fresh and dry weight in iodine deficiency, and the iodine content markedly decreases. They found that 1 to 2 micrograms of iodine per rat daily will prevent thyroid enlargement.

Chesney *et al.*[181] reported goiter in rabbits fed a diet consisting solely of cabbage. These goiters were associated with a lowering of the metabolic rate, and feeding iodine raised the metabolic rate and prevented thyroid hyperplasia. Baumann *et al.*[182] confirmed the findings and concluded that cabbage contains a goiterogenic substance which acts by depleting the thyroxin store of the thyroid gland.

DIETS FOR LABORATORY ANIMALS

The most successful results in raising laboratory animals are obtained in practice when a variety of good quality natural food materials are available. Because of the wide distribution of most nutrients in common foods, there are innumerable combinations of foodstuffs which would adequately supply the nutritional needs of laboratory animals. A few such combinations of natural food materials will be suggested.

It has been possible to devise diets of highly purified ingredients which permit excellent growth in rats and mice. While instances of reproduction have also been recorded, animals have not been maintained for many generations upon purified diets as a final test of their completeness. Furthermore, upon such diets, only limited success has been achieved with rabbits and guinea pigs.

[177] J. Biol. Chem. **94:** 117. 1931–2.
[178] J. Biol. Chem. **109:** 347. 1935.
[179] Am. J. Physiol. **119:** 768. 1937; **124:** 750. 1938.
[180] J. Nutr. **6:** 325. 1933; **15:** 539. 1938.
[181] Bull. Johns Hopkins Hosp. **43:** 261. 1928.
[182] Proc. Soc. Exp. Biol. Med. **28:** 1017. 1931.

In 1930, Maynard[183] reported a ready-mixed, commercially available stock diet for rats which gave satisfactory results during growth and reproduction. The mixture was composed of linseed oil meal, 300 pounds; ground malted barley, 200; wheat red dog flour, 440; dried skim milk, 300; oat flour, 300; yellow corn meal, 400; steamed bone meal, 20; ground limestone, 20; and salt, 20 pounds. Extra vitamins A and D were supplied twice a week but no green feed or other supplements were fed. This same mixture has served as a stock diet in the Cornell rat colony until the present time (October, 1944) and it is considered to be an adequate diet. It has also been successfully used during many generations for the mouse and hamster, and for the rabbit, when fed along with hay; and for the guinea pig, when extra roughage and vitamin C was supplied. Recent tests in our laboratory and a comparison of the growth and reproduction of other rat colonies (Smith[184]) clearly demonstrated that the calf meal, as the only food, is far from optimum for rats.

Mendel and Hubbell[185] fed their rats, in addition to the calf meal described above, other supplements which have improved growth and reproduction. Nursing mothers and young rats under 6 weeks of age received a "paste food" consisting of casein 25 per cent, whole milk powder 25, wheat embryo 20, and lard 30 per cent. The calf meal and the paste food were both supplied *ad libitum*. Each rat also received 1 gm. of dried yeast daily except Sunday, and those without paste food were given 3 gm. wheat embryo per week. No "green" food of any sort was used. The comparative rates of growth of rats fed calf meal only and those fed extra supplements are shown in FIGURE 1. Data on reproductive efficiency are presented in TABLE 2.

The marked improvement in the rate of growth, the larger adult size and the superior reproduction are most probably the result of dietary improvement due to the paste food and yeast.

Sherman and associates[186] have shown that a diet made up of about five parts whole wheat and one part whole milk powder with 2 per cent of salt is adequate for growth and reproduction of rats during many generations. Nevertheless, such a diet is not optimum, for they have observed that supplements which furnish extra calcium, riboflavin or protein resulted in more rapid growth, earlier maturity, longer reproductive life, larger adult size and longer life. Adding extra fat resulted in a slightly slower growth rate but also an increased length of life.

[183] Science 71: 192. 1930.
[184] U. S. D. A. and Cornell University. Unpublished data.
[185] J. Nutr. 10: 557. 1935.
[186] J. Nutr. 14: 609. 1937; 16: 603. 1938.

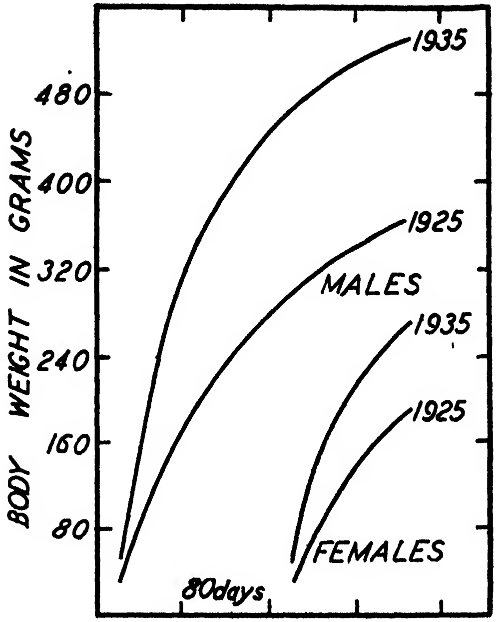


FIGURE 1
Growth of rats on calf meal (1925) and on calf meal plus the "paste food" supplement

TABLE 2
REPRODUCTIVE PERFORMANCE UNDER DIFFERENT DIETARY CONDITIONS

	Fertile matings	No. in litter	Birth wt.	Weaned	Weaning Wt.		Av. daily gain to 100 days	
					Males	Females	Males	Females
	%		gm.	%	gm.	gm.	gm.	gm.
1919 to 1925	68	6.4	—	76	31	30	2 1	1.6
1925 to 1935	93	9.6	5.8	90	48	47	4.0	2.5

That appreciable fat in the diet of rats is essential for optimum lactation performance is clearly demonstrated by the studies of Maynard *et al.*[187] Adding 10 to 20 per cent of fat to the diet of females resulted in a marked increase in the growth-rate of the young during the suckling period. When fat was added to calf meal or to a low-fat synthetic

[187] J. Nutr. 23: 385. 1942, 28: 81. 1944.

diet to increase the amount to 10 per cent, young rats gained, as an average, 1.5 grams during the first 17 days, as compared with about 1.0 gram without the extra fat. At weaning (21 days), young on high-fat weigh 40 to 45 grams as compared with 25 to 30 grams for young fed the low-fat diets.

The importance of this rapid gain in weight for rats to be used in assay studies has been stressed by Anderson and Smith.[188] These authors have reviewed the reports since 1906 showing the average weights of the albino rat. During this time, a marked increase in the rate of growth has occurred. They express the view that improved feeding practice has been a major cause of the more rapid growth rate. Selected data which illustrate the increased growth rate of albino rats are shown in TABLE 3.

TABLE 3
THE AVERAGE GAIN IN WEIGHT OF ALBINO MALE RATS

Date and Reference	Body weight at 100 days of age	Average daily gain
	gm.	*gm.*
1906 (Donaldson *et al.* [189])	165	1.65
1915 (King [190])	200	2.0
1928 (Smith and Bing [191])	315	3.2
1935 (Mendel and Hubbell [185])	—	4.0

In a recent survey Smith[184] has compared the rate of growth and the reproductive efficiency of a number of rat colonies. The litter size and the body weights of males from these colonies are shown in TABLE 4.

The striking difference in the performance of these rat colonies is extremely interesting. It is not possible to indicate with certainty the relative importance of genetic differences and of feeding and management practices as causes of these variations. When it is recalled, however, that the Connecticut colony (Mendel and Hubbell[185]) increased in breeding efficiency and rate of growth from the level of the other colonies shown in table 4 to its present high rating (see table 2), at least, largely as the result of a dietary change, it appears worth while to consider the diets used in these colonies. A comparison of the diets is shown in TABLE 5.

[188] Am. J. Physiol. 100: 511. 1932.
[189] Boas Anniversary Volume 5. 1909. New York.
[190] Anat. Record 9: 751. 1915.
[191] J. Nutr. 1: 179. 1928.

TABLE 4

LITTER SIZE AND BODY WEIGHT OF DIFFERENT RAT COLONIES

Colony	Young in 1st litter	Weight of males		
		At weaning	At 100 days	Maximum
	No.	*gm.*	*gm.*	*gm.*
Columbia	6.8	52.6[2]	274	388
Connecticut	9.5	51.8[1]	358	531
Cornell	6.2	29.7[1]	234	361
Penn State				
Pied	7.4	—	—	—
Albino	7.9	—	—	—
Wild	7.4	—	—	—
U. S. D. A.	6.4	65 6[2]	276	434

[1] 21 days of age.
[2] 28 days of age.

TABLE 5

A COMPARISON OF DIETS FED TO DIFFERENT RAT COLONIES

	Columbia	Connecticut[1]	Cornell	Penn State[2]	U.S.D.A.[3]
Wheat	5	—	—	18	33
Yellow corn	—	20	20	67	33
Malted barley	—	10	10	—	—
Wheat red dog	—	22	22	—	—
Oats or oat flour	—	15	15	2	10
Skim milk powder	—	13	13	—	10
Whole milk powder	1	—	—	2	—
Soluble blood flour	—	2	2	—	—
Wheat bran	—	—	—	1	—
Meat scrap	—	—	—	2	—
Alfalfa leaf meal	—	—	—	2	3
Casein	—	—	—	—	4.5
Lard	—	—	—	—	5.5
Salt	1.6	1	1	0.4	0.5
$CaCO_3$	—	—	—	0.6	0.5
Ferric citrate	—	—	—	0.04	—

[1] Nursing dams and young to 6 weeks of age are also given a "paste food" consisting of casein, 25; whole milk powder, 25, wheat embryo, 20; lard, 30. Each rat is given 1 gm. of yeast daily and 3 gm. wheat embryo, each week, when they receive no "paste food."
[2] Each rat receives about 10 cc. whole milk daily and, except nursing dams, 5 cc. tomato juice 3 times weekly.
[3] The rats are fed whole milk 3 times weekly.

Preliminary studies now in progress at Cornell suggest that much of the difference between the five rat colonies studied by Smith[184] probably can be attributed to dietary differences.

Excellent growth can be achieved when rats are fed diets made up of highly purified food ingredients and supplemented with crystalline vita-

mins. Reproduction upon such a dietary regime has been less successful. Vinson and Cerecedo[192] have recently reviewed earlier studies and reported purified diets which maintained growth, reproduction and lactation through four generations. Selected data on reproduction and lactation presented by these authors for Wistar strain of rats are shown in TABLE 6.

TABLE 6
REPRODUCTION ON HIGHLY PURIFIED DIETS

Generation	Litter size	Average weaning wt.
	no.	*gm.*
Parent	6.7	38.3
F_1	5.3	33.8
F_2	4.0	34.2
F_3	4.0	29.6
Controls on Purina dog chow	7.0	34.5

While growth was rapid upon the diets employed, lactation was poor. The authors suggest that some substance (lactagogue), causing increased milk production, is lacking in the purified diets. Such a sub-substance seems to be present in yeast. The data also show that reproduction is less efficient than upon the control diet.

Morse and Schmidt[193] have shown that nitrogen is lost during gestation on a synthetic diet, but stored on a diet of natural feeds. Nitrogen losses during lactation were greatly increased on the synthetic diet.

THE COTTON RAT will probably find more use as a laboratory animal and, therefore, the nutritive requirements of this animal are of interest. Jungeblut[194] has shown that this species does not require a dietary source of ascorbic acid. McIntire, Schweigert and Elvehjem[195] list the B vitamin requirements in micrograms per 100 grams of diet as follows: thiamine 150, pyridoxine 100, pantothenic acid 800, riboflavin between 80 and 300, nicotinic acid less than 2,500, inositol less than 100,000, choline less than 100,000, and other factors present in liver extract.

THE HAMSTER: Coopermann, Waisman and Elvehjem[196] recently studied the nutritional requirements of the golden hamster. They fed a basal diet of sucrose 72, casein 18, salts 5, corn oil 2, cod liver oil 2, and wheat germ oil 1, with 1 drop of haliver oil every two weeks. Sat-

192 Arch Biochem. 3: 389. 1944.
193 Proc. Soc. Exp. Biol. Med. 56: 57. 1944.
194 J. Nutr. 20: 427. 1940.
195 J. Nutr. 27: 1, 1944.
196 Proc. Soc. Expt. Biol. Med. 52: 250. 1943.

isfactory growth and instances of reproduction were observed when thiamine, riboflavin, calcium pantothenate, pyridoxine, sodium para-aminobenzoate, inositol, choline, and biotin were supplied. Their data support the view that nicotinic acid is not essential. Routh and Houchin[197] had earlier reported that the hamster needed nicotinic acid. Vitamin C is not a dietary essential. On a biotin-deficient diet Cooperman *et al.*, observed a characteristic dermatitis at the corners of the mouth.

A diet of natural food materials which is satisfactory for the albino rat will probably also meet the needs of the cotton rat and the hamster.

Rabbits and guinea pigs: Hogan and Ritchie,[198] in attempts to rear rabbits and guinea pigs on purified diets, found they failed to grow and survive on diets which were satisfactory for rats. Madsen *et al.*,[48] and Davis *et al.*[199] obtained fair growth on certain purified rations. Shen[200] observed approximately normal growth of guinea pigs and survival up to two years on a diet of casein 15, starch 35.5, sucrose 15, cellulose 15, agar 5, salts 4.5, cottonseed oil 5, and yeast 6, supplemented with irradiated yeast, carotene and tomato juice. A number of the females bred, but they generally aborted or delivered dead young.

Kohler *et al.*[201] reported a "grass juice factor" which is essential for growth of guinea pigs. Cannon and Emerson[202] obtained data which they interpreted as a confirmation of the existence of this dietary essential.

Sober *et al.*,[203] using guinea pigs, obtained most rapid growth and longest survival on a purified diet containing sucrose 62, casein 30, salts 4, corn oil 4, thiamine 250 mg., pyridoxine 250, riboflavin 350, pantothenic acid 1000, niacin 5 mg. and choline 400 mg. The diet was supplemented with yeast 16 per cent, dried grass 16 per cent, and 20 ml. whole milk. Each animal was given 10 mg. ascorbic acid daily and 1.0 mg. a-tocopherol acetate and 4 drops haliver oil weekly. On this diet the animals survived at least 7 weeks and gained an average of 3.8 gm. per day. Hogan and Hamilton[204] found guinea pigs would not grow or survive long on purified diets containing only the known crystalline vitamins. Dried yeast and a water extract of liver contain the unidentified essentials. These authors doubt the existence of an essential "grass juice factor" for the guinea pig. Reproduction was obtained

[197] Federation Proc. **1**: 191. 1942.
[198] Missouri Agr. Exp. Sta. Bull. **219**. 1934; **370**: 23. 1936.
[199] Cornell Agr. Exp. Sta. Memoir **217**. 1938.
[200] Thesis, Cornell University, Ithaca, N. Y. 1939.
[201] J. Nutr. **15**: 445. 1938.
[202] J. Nutr. **18**: 155. 1939.
[203] J. Nutr. **24**: 503. 1942.
[204] J. Nutr. **23**: 533. 1942.

in rabbits when extracts from yeast and liver were fed. Wooley[205] and Kuiken *et al.*[206] reported evidence suggesting that guinea pigs require three more dietary factors than do rats and mice. Two of these factors are furnished by linseed oil meal.

Mannering *et al.*[207] confirm the presence in linseed oil meal of two factors which are essential for guinea pigs. They show data supporting the view that essential factors are also furnished by solubilized liver powder and grass juice powder.

It has not been conclusively established that rabbits and guinea pigs require all known members of the vitamin B complex, but it appears certain that they need two or three vitamins which are not now available in crystalline form. Guinea pigs need vitamin C to prevent scurvy while rabbits appear to be able to synthesize ascorbic acid. In selecting stock rations, it is imperative that some green feed be constantly available for guinea pigs to supply ascorbic acid.

Bischoff and Sansum[208] maintained rabbits for two years on alfalfa. When barley was the only food, rabbits developed fatty livers and nephritis. Hogan and Ritchie[198] fed their stock rabbits whole oats and alfalfa hay, plus about 75 cc. of whole milk of lactating females.

At Cornell, rabbits and guinea pigs have been successfully maintained on the calf meal (TABLE 5) fed to rats and hamsters. Good quality fine stemmed legume or mixed hay is also fed *ad libitum* and guinea pigs receive green feed (grass, carrots, cabbage, lettuce or sprouted grains) two or three times weekly. Good results have also been obtained with a number of dry dog foods which are generally available. A very simple mixture of whole wheat 2, whole oats 2, and linseed oil meal 1 (pelleted or pea size coke), as suggested by Templeton *et al.*,[209] has given fully as satisfactory results as the more complicated feeds. With the latter mixture, it is advisable to supply a mineralized salt block in addition to the hay. Slanetz[210] has recently studied the nutritive quality of several commercial feeds especially prepared for laboratory animals.

---

[205] J. Biol. Chem. 148: 679. 1942.
[206] J. Nutr. 27: 395. 1944.
[207] J. Biol. Chem. 151: 101. 1943.
[208] J. Nutr. 5: 403. 1932.
[209] U. S. D. I. Conservation Bull. 25. 1942.
[210] Am. J. Vet. Res. 4: 182. 1943.

DISCUSSION OF THE PAPER

Dr. George R. Cowgill (*Professor of Nutrition, Yale University*):

Dr. Loosli's paper surveys very well the individual dietary factors important for successful feeding of various species of animals. A few comments can be made regarding some of these factors.

PROTEINS AND AMINO ACIDS—Some interesting observations have been made of animals subsisting on diets containing less than the optimal amount of protein. Students of hormone problems have a useful tool in the young female rat forced to grow while subsisting on a diet containing about 4 to 6 per cent protein, instead of the 12 to 18 or more per cent used by most investigators. In our laboratory at Yale some years ago, Drs. Orten and Smith (1937)[1] observed that young female rats fed diets containing 3.5 per cent lactalbumin grew somewhat, but definitely failed to mature sexually. Many interesting observations were made on these animals. The gonads were able to respond to an injection of pregnant mare's serum, proving that the condition was one of lack of gonadotropic stimulation rather than actual injury by the dietary regime. We may speculate on the possibility that this represented a failure of the pituitary gland to grow properly in the face of a limited supply of certain essential amino acids. It is known that the hormones elaborated by the pituitary gland are protein in nature and that these proteins contain at least some of the amino acids known to be essential for nutrition of the rat. We may suppose that, in the face of a limited supply of certain essential amino acids, the growing pituitary gland is forced to compete with other structures of the body with the result that it fails to make the customary growth, and, as a result of this, many other structures which depend on the pituitary gland fail to grow properly. This would explain the sexual immaturity observed in these animals. We see in observations of this sort an illustration of how the diet can be an important factor for students of the physiology of sex to consider in their researches.

CALORIES—Dr. Loosli did not mention the fact that, broadly speaking, animals eat calories, although this is implied in much of his discussion It is this phenomenon of "eating calories" that makes it very necessary for us, when performing feeding experiments and endeavoring to make quantitative observations, to see that the food intake is measured, if possible. Since an animal eats enough food to meet the energy requirement, it will eat considerably more of a low-fat diet in order to secure those needed calories than it will of a high-fat diet. This indicates the importance in feeding experiments of knowing not only the percentage composition of the diet but the energy value as well. Incidentally, I might add that it is as a rule more difficult to secure accurate intake data when feeding a low-fat diet because of its powdery nature, than it is when feeding a high-fat diet which is likely to be paste-like and more solid and therefore more difficult for the experimental animals to scatter. When feeding a powdery diet, various ingenious devices have been employed to keep the animals from scattering the food.

FAT—Dr. Loosli ably summarized the literature on the importance of fat in the diet. A word or two might be said regarding some practical aspects of this matter. Most dry animal feeds that are not marketed in a sealed container—a can for example—are very low in fat. This is due to the tendency of high-fat mixtures to turn rancid and therefore to spoil before they can be used by the purchaser.

In our own feeding experiments with dogs, we have become impressed with the importance of knowing, for each animal, what we call the Nutritive Index. By the Nutritive Index, we mean the ratio of the cube root of the body weight to the length of the animal, *i.e.*, $\frac{\sqrt[3]{\mathrm{WgHT}}}{\mathrm{LA}}$ (Cowgill and Drabkin, 1927[2]; Cowgill, 1928[3]). We have measured this Nutritive Index in many dogs and satisfied ourselves that the maximum value is about 0.34, when weight is expressed in grams and length in centimeters and the weight is the maximum possible for the unit length, for example, the value for an extremely obese dog. This maximum value, namely 0.34, may be compared with the value yielded by the animal in question. If this animal is not as completely filled out in all three dimensions, so to speak, for its

[1] **Orten, Aline U., & A. H. Smith.** The effect of dietary protein on endocrine function and on the blood picture of female rats. Am. J. Physiol. **119**: 381. 1937.

[2] **Cowgill, G. R., & D. L. Drabkin.** Determination of a formula for the surface area of the dog together with a consideration of formulae available for other species. 1927.

[3] **Cowgill, G. R.** The energy factor in relation to food intake: experiments on the dog. Am. J. Physiol. **85**: 45. 1928.

length, then the value will be less than 0.34. A group of dogs was allowed to have all that it wished of diets of known composition. After a period, each animal put itself into such a nutritive state that the value of this ratio in every case was very close to 0.30. We, therefore, take 0.30 as the value representing a good state of nutrition that the dog will endeavor to reach if allowed all it wishes of good food. If a dog is fasted sufficiently, the value obviously drops considerably below 0.30.

Some years ago, we had occasion to use our Nutritive Index in an interesting way on dogs being used by Dr. Hansen for studies of fat metabolism. Interesting values for blood lipids had been secured. Dr. Hansen had tried to make his animals more or less equal at the beginning of the study by fasting each of them for a short period. However, some of the animals were in excellent nutritive condition when such a fast began, whereas others were not so well nourished. Obviously, a short fast does not seriously affect an animal that has a large fat reserve; a dog with little fat reserve, when subjected to a fast, quickly begins to burn his own body tissue. Many of Dr. Hansen's data, apparently, could not be explained. When, however, he determined the Nutritive Index of each animal, it was perfectly obvious what had happened and it was clear that the dogs were not the same just because of the short-fast.

We have consistently used this principle of Nutritive Index in our work with dogs. Whenever dogs are put on metabolism experiments, where they are expected to eat all of the food offered and to be similar with respect to what we like to regard as a "known nutritive state," the Nutritive Index is determined and then the animal fasted or liberally fed until the Nutritive Index approximates a value slightly below 0.30. Then the particular experiment of interest is started. By this procedure, we have been able to eliminate a large amount of variation in metabolic data yielded by our dogs. Perhaps, this has some application to other species. So far as I know, this point has not been tested.

Vitamin A and Carotene—Dr. Loosli has pointed out the importance of vitamin A in nutrition and its relation to the provitamin carotene. One might call attention to the fact that different species show different degrees of utilization of carotene in studies where vitamin A is the variable of interest. This fact should be taken into account.

As Dr. Loosli pointed out, vitamin A deficiency is characterized by changes in the epithelial tissues. This has a bearing on the problem of infections in the colony. Some years ago, with my colleague, Dr. H. M. Zimmerman, studies were made on vitamin A deficient rats to determine whether certain lesions develop in the nervous system. Many groups of animals were started, but they failed to survive long enough for such lesions to develop. We, therefore, isolated the animals in a private room and instructed the assistant who took care of the animals to use what amounted to an aseptic technique as far as possible. As a result, we had perfect success. The animals did not die of infections, but developed a very definite and characteristic paralysis. Treatment with large doses of carotene resulted in restoration of growth, a good hair coat, a return to a general nutritive condition, but no cure of the paralysis. This experience taught us the importance, in any vitamin A deficiency studies, of protecting the animals against ordinary infections. Presumably, the fall in resistance to infection is related to the keratinizing of the epithelial cells lining the upper respiratory passages.

Vitamin E—This vitamin has acquired a significance in nutrition somewhat different from that originally attributed to it. We now know the chemical structure of vitamin E and some of its important chemical properties. We know that it is a very valuable anti-oxidant. According to the most recent work of Hickman and associates (1944),[4] it can serve to protect many other factors, notably vitamin A, carotene, and even ascorbic acid from oxidation. Other workers have shown that batches of experimental diets made up in any quantity for laboratory work, if allowed to become rancid, will lose their content of vitamin E. In experimental

[4] **Hickman, K. C., M. W. Kaley, & P. L. Harris.** Covitamin studies. I. The sparing action of natural tocopherol concentrates on vitamin A. J. Biol. Chem. 152: 303. 1944; II. The sparing action of natural tocopherol concentrates on carotene. J. Biol. Chem. 152: 313. 1944; III. The sparing equivalence of the tocopherols and mode of action. J. Biol. Chem. 152: 321. 1944.

work, therefore, it is important to take measures to avoid the development of such rancidity. There is a very practical and important question here to which the investigator must give careful attention, if he is to be sure that any vitamin E added to the food mixture will still be there when the animal actually eats the food.

B-COMPLEX FACTORS—The latest interesting findings bearing on these factors, that might be commented upon here, relate to the production of several of the B vitamins in the intestine by bacteria. This helps to explain the observations made many years ago that animals allowed access to their feces can get along with smaller amounts of B-complex vitamins in their food. We now know that the feces contain many of these factors and also that the amounts present in the feces can be influenced by the composition of the diet. It is for this reason, among others, that students of the B-complex problems almost invariably house their animals in raised-screen bottom cages, through which the feces can escape from the animal. Many years ago, we had occasion to measure the amount of vitamin B needed by young rats housed in the old style cage, as compared with that required when the animals were in cages provided with screen bottoms. The requirement for B proved to be roughly twice that when the old-style cage was used. The bearing of this on many feeding experiments seems obvious. One might also cite much literature showing that the carbohydrate in the diet can affect the amounts of various members of the B vitamins present in the feces. Apparently, the carbohydrate which has the least effect and is, therefore, an important one to use in experimental studies of this general problem is sucrose. When starch, or dextrin, or lactose is used, analyses of the feces show interesting differences in amounts of various B-complex factors present.

Mention might also be made of the fact that the ingestion of raw egg-white results in a binding of the biotin present in the food as well as that produced by the intestinal bacteria, with the result that the feeding of a diet containing an appreciable amount of raw egg-white can result in the signs of biotin deficiency. Also, many sulfa drugs have been found to inhibit the growth of certain intestinal bacteria. By the use of such drugs, it has even been proven possible to produce in the animal, under certain conditions, the characteristic signs of deficiencies of certain of the B-complex factors.

VITAMIN C—Dr. Loosli has pointed out that not all species require this factor. Man, the anthropoids and the guinea pig are species that are unable to synthesize ascorbic acid sufficiently to meet their requirements Time will not permit a discussion of many interesting phases of vitamin C functions in nutrition. It does seem pertinent, however, to comment here on the use of guinea pigs in the laboratory. I have been surprised at the number of "practical" animal raisers who believe the old claim that a guinea pig does not need to be provided with water, as such, "because this animal eats such large amounts of water-rich leaves—lettuce, etc.—enough water can be obtained from such materials." One guinea pig breeder said that, when he put water in the cage, the results were very bad because the animals would get wet. It has been our experience that, when the cage is provided with a proper drinking fountain, so that the animal can get its water without getting a bath, so to speak, they drink really large amounts. This indicates that they certainly have a water requirement that should not be neglected and which can be easily taken care of by suitable cage arrangement. This merely confirms what any student of nutrition can quickly prove by simple calculations and considerations of metabolism. I doubt whether any competent research worker with guinea pigs neglects to provide water in this fashion. Apparently, this false notion prevails almost entirely among the so-called "practical" animal raisers.

MINERAL NUTRIENTS—In considering this group of dietary factors it seemed appropriate to mention the work of Richter (1942–43)[5] at Johns Hopkins University. Dr. Richter, as you know, has been interested in what experimental animals choose to eat when allowed to make many choices of one kind or another. A study of such responses leads to the conclusion that any marked and unusual

[5] **Richter, C. P.** Total self regulatory functions in animals and human beings. The Harvey Lectures Series 38: 63–103. 1942–1943.

"appetite" means a special requirement for that factor which the animal is endeavoring to meet. People who operate animal colonies and who desire to study this kind of a problem will find much of profit in the cage devices and experimental arrangements devised by Dr. Richter.

General Comments—It might be pertinent here to make some reference to all-glass cages which many laboratories have used to advantage, particularly when studying problems involving trace elements of one kind or another. Our laboratory has been interested in problems relating to lead. We have used an all-glass cage for young rats and, by this device, together with the feeding of diets extremely low in lead, have been able to get animals with a lead content lower than any reported in the literature. We have even used special stainless steel cages for similar experiments on puppies and with equally satisfactory results.

Not all workers are aware of the importance of checking the effect of some simple laboratory manipulation of a food or the diet on its chemical composition. One illustration of what I have in mind will suffice. Studies have been made of the mineral composition of cereal grain and other plants grown under special conditions. It has been found that, if the wheat or other material that has been carefully grown under carefully controlled conditions is ground up in an ordinary mill made of iron, just to facilitate the chemical analysis, very appreciable amounts of iron can be added to the material being analysed. Erroneous conclusions that have been published regarding the iron content of such materials have been readily explained in this way. One wonders whether this sort of error may not have crept into other food analyses reported in the literature. When careful analyses are being made of substances present in very small amounts, one should not ignore the possible effects of various simple laboratory procedures or special manipulations of the material.

Dr. M. H. Ross (*Biochemical Foundation, Newark, Del.*):

We have had to feed large numbers of dogs and lately have been having trouble getting meat. Perhaps some one has some suggestions to offer.

Dr. Ellis J. Robinson (*American Cyanamid Co., Stamford, Conn.*):

We have been maintaining 40 to 50 dogs and have had a good supply of meat up to the present.

Dr. C. A. Slanetz (*College of Physicians and Surgeons, Columbia University, New York, N. Y.*):

During the last 8 years we have had 200 dogs and have fed them common dog meal. I would suggest you get in touch with the American Medical Association.

Dr. B. J. McGroarty (*Sharpe and Dohme, Glenolden, Pa.*):

We have been very well supplied with meat for our dogs used in work on rabies.

Dr. Ross:

Manufacturers change the formula for the constituents of the food according to what the market has available. However, the protein and carbohydrate content of the formula may remain the same. Many of the products are imported from Canada.

Dr. Robinson:

We have been following the hematology of rats fed the same common diet and for over five years the whole pathological picture has remained very much the same. One needs to know the additional weight of the animal when starting an experiment and how the amount of food will change the requirements.

Dr. Cowgill:

I am very much surprised that people think meat is essential in the diet of dogs. At Yale, dogs have been fed, for many years, on highly purified diets without any meat and good results have been obtained.

Dr. Clarence A. Mills (*College of Medicine, University of Cincinnati, Cincinnati, O.*):

I should like to confirm Doctor Cowgill's remark on maintaining dogs on a simple diet. We have maintained dogs on a synthetic diet and they take it as well as meat and often prefer it. This is not true of cats. Cats are harder to keep than other animals. One must find a diet that is palatable to the cat as cats will often go on a hunger strike. I would like to hear more discussion on the subject.

Dr. T. F. Zucker (*College of Physicians and Surgeons, Columbia University, New York, N. Y.*):

There is not much information on work with cats but it has been found that the animals' diet can be improved by adding miscellaneous items such as B complex (wheat germ). Carotene is an ineffective element in the diet of the cat. Early work on the diet of the cat has shown that diet deficiencies are directly related to the growth and stature of the cat.

INFECTIOUS DISEASES OF LABORATORY ANIMALS

By Herbert L. Ratcliffe

Penrose Research Laboratory, Zoological Society of Philadelphia and Department of Pathology, University of Pennsylvania.

Whatever the field of investigation, or the immediate problem to be studied, one may say that the experimental biologist more often than not has hope rather than assurance that his animal subjects are healthy. When one is concerned only with short-term studies, perhaps the source and care of experimental animals need not be considered so completely. Yet, certainly, any experiment demands that variables be reduced as much as is possible, which, in turn, demands standardized animals maintained under standardized conditions. This goal may not be completely attainable, animals being what they are. One obstacle is the considerable number of infectious agents to which laboratory animals are susceptible.

Under natural conditions, the species from which laboratory animals are derived are subject to many of the same infections that occur in captivity, yet they survive and multiply. One wonders, at times, whether or not efforts have been misdirected. Possibly, colonies of laboratory animals should be fully exposed to infections, until, by selection, highly resistant strains are obtained. Certainly, present-day efforts to exclude infections develop highly susceptible races.

Nevertheless, the chief concern of this review will be the more important infectious disease of the following species of animals: rats, mice, guinea pigs, rabbits, dogs, cats, monkeys, and canaries. To a large extent, housing and care of colonies of these animals are modified by the necessity of guarding against infections. Hence, these phases of the topic will be considered only as they apply to control of contagious disease.

Location of animal-rooms and their construction, the size and arrangement of cages, the temperature at which quarters are kept, and hygienic considerations have varied widely as different laboratories have sought to meet the needs of one or another species. Certainly, considerable variation is possible without neglecting essentials.

This much is certain, animal-rooms should be proof against wild mice and rats, and against insects. They should be adequately lighted, ventilated and heated. Walls and floors should withstand thorough washing, with satisfactory drains to allow rapid drying. Tables, cage-

racks and other installations should be so constructed that they do not interfere with cleaning.

When colonies are to occupy more than one room, each room should be so constructed that it may be completely isolated. Then, outbreaks of infectious disease may be checked, sick animals isolated and new arrivals quarantined. When animals are used in experiments involving infectious agents, they should, of course, be completely isolated from other susceptible stock and all possible sources of contact controlled.

Construction and arrangement of cages, obviously, must be governed by the species of animal and purpose for which it is intended. Whatever the type and size of the cages, they should be large enough to permit freedom of movement, allow ready access to animals without undue danger of escape, be easily cleaned, and withstand repeated sterilization. Laboratory supply houses offer many types of cages, few of which meet these requirements

Rodents, rabbits, and monkeys should be provided with some form of bedding. Straw, saw-dust, shavings, and grain-wastes are commonly employed. Whatever the material used, it must be regarded as a vehicle by which insects, arachnids, and other pests may be introduced into the colony. Mite infection of the skin of mink has been traced to contaminated rice hulls[1] Saw-dust, wood-shavings, or cut-paper may contain "bed-bugs" (*Cimex*), or dog and cat feces. "Bed-bugs" suck blood and may cause considerable debility. Ova of tapeworms contained in cat or dog feces readily infect rabbits, rats, and mice. Therefore, bedding should be sterilized and stored in safe containers. How frequently bedding material need be changed will, of course, depend upon the species of animal, and its space allowance.

As in all animals, recognition of disease in laboratory animals requires some experience. One cannot stress too strongly the fact that evidences of illness are most quickly recognized by those who know the appearance and behavior of normal animals. Considerable time may be well spent simply in observing normal animals, for colony caretakers may sometimes overlook abnormalities which, if recognized early, often allow more effective control of epidemic disease.

With few exceptions, the onset of acute infectious disease is marked by inactivity, loss of appetite, rough coat, and dull, partly closed eyes. These signs merely warn of possible danger, but are no guide to its specific nature. So, too, others changes that may be noted in the animal, either before or after death, rarely identify the disease.

[1] Bushnell, F. R., & O. J. Kumon. Fur Trade J. 21: 5–6 Canada 1944.

Usually, specific diagnosis requires identification of the pathogenic organism by microscopic examination of appropriate material, by bacteriologic techniques, or by animal inoculation.[2]

Descriptions of diseases of laboratory animals are widely scattered and often confused. Fortunately for me, this is not the first effort to assemble and digest this information. I am especially indebted to K. F. Meyer,[3] whose review of the infectious disease of rabbits, guinea pigs, rats and mice has simplified my task immeasurably, and to J. H. Dingle,[4] whose more recent discussion of the infectious disease of mice should be carefully studied by those interested in this species. Both of these papers contain vivid descriptions of disease processes and clear directions for diagnosis and control, and both have extensive bibliographies. The present discussion duplicates much that is contained in these papers, and omits consideration of many of the less common diseases that are covered by these authors.

The various species of laboratory animals that are considered in this review are susceptible to a host of infections. Some of these are specific to the species, others have a much wider range. The following descriptions of disease are arranged in the order of their apparent significance.

BACTERIAL DISEASE

(Ref. 5, 6, 7, 8, 9)

1. **Salmonellosis.** Bacteria of the *Salmonella* group are potentially, if not actually, the most important causes of disease among laboratory animals, especially the small rodents. In most instances, outbreaks are caused by two species, *S. typhimurium* or *S. enteritidis,* singly, together, or in combination with other organisms of this group. Both species are as widely distributed as their natural hosts, wild mice and rats, and are as ever-present. Guinea pigs, rats, mice, canaries, and other birds are highly susceptible. Rabbits may be infected by contact with mice or rats that are carriers, and, on occasion, the organisms

[2] **Topley, W. W. C., & G. S. Wilson.** The Principles of Bacteriology and Immunity. 2nd. ed. William Wood & Co. Baltimore. 1937.
[3] **Meyer, K. F.** The Newer Knowledge of Bacteriology and Immunology: 607–638. Univ. of Chicago Press 1928.
[4] **Dingle, J. H.** Biology of the Laboratory Mouse: 380–474. Blakiston Co., Philadelphia. 1941.
[5] **Lovell, R.** Vet. Rec. **12**: 1052–1065. 1932.
[6] **Hagan, W. A.** The Infectious Disease of Domestic Animals. Comstock Publishing Co., Ithaca, N. Y. 1943.
[7] **Jaffe, R.** Anatomie und Pathologie der Spontanerkrankungen der kleinen Laboratoriumstiere. Springer, Berlin. 1931.
[8] **Beaudette, F. R.** J. Am. Vet. Med. Assoc. **68**: 642–643. 1926.
[9] **Ratcliffe, H. L.** The Rat in Laboratory Investigations: 456–471. Lippincott Co., Philadelphia. 1942.

have been isolated from cases of acute intestinal disease of dogs and monkeys.[5]

Both wild mice and rats, as well as laboratory stocks of these animals, are common carriers of *S. typhimurium* and *S. enteritidis,* but female guinea pigs may fill the role also; in which case, they serve mainly as a continued source of infection for their colony. Epidemics of salmonellosis in guinea pigs are commonly associated with the breeding season.

The usual route of *Salmonella* infection is by mouth. Contaminated food is the common vehicle. Organisms may enter the colony in food to which wild mice and rats have had access, or in new stock that is added without adequate quarantine, and again the colony may maintain its own sources of infection in mice, rats, or guinea pigs which carry inapparent disease.

When introduced into a susceptible stock, *Salmonella* infections follow the pattern of any acute infectious disease. Some animals die within 48 to 72 hours after exposure, others after several days or weeks. Some recover after more or less prolonged disease, and a few have unapparent infections. Mortality rates vary, but often are higher than 75 per cent of the colony.

Salmonella infections are not distinctive, either *ante mortem* or at autopsy, but must be differentiated from other diseases by bacteriological examinations of the feces before death, or of the liver and spleen after death. After an incubation period of three or six days, signs of acute disease make their appearance. Mice and rats usually develop conjunctivitis and diarrhea, and diarrhea may occur in other species also; but, more commonly, this disease is subacute or chronic, with signs of illness in keeping with the course of the disease. In guinea pig colonies, the disease may make its appearance only during the breeding season, many animals aborting and dying of purulent endometritis.

In animals dying of the more acute types of salmonellosis, one finds only congestion of the viscera, slight enlargement of the liver, spleen and mesenteric lymph nodes, and early catarrhal inflammation of the intestine. When the disease has progressed more slowly, loss of weight may be pronounced. In any event, the most significant pathological changes are limited to the abdominal viscera. The liver, spleen, and mesenteric lymph nodes are greatly enlarged and contain many small pale grey areas of necrosis. Or one may find larger, yellowish abscess-like areas in the spleen and lymph nodes of the mesen-

tery. Usually, the intestinal wall is thickened and injected, and its lymphatic tissue is hyperplastic, and contains yellowish foci of necrosis. Ulcerations of the mucosa of the lower ileum and caecum may be found also. These lesions are fairly uniform in all species that are susceptible to *Salmonella* infection, but the degree to which they are developed depends largely on the duration of the disease. In addition to these changes in the abdominal viscera, it is said that guinea pigs also develop endometrial infection and pneumonia, lesions that can follow infection by many other organisms.

Prevention and control of *Salmonella* infections is difficult. The organisms are so wide spread and may be introduced so easily, that only a carefully developed routine will maintain colonies of mice, rats, guinea pigs, and canaries free of infection. 'Once infection is established in a colony, destroying the herd and obtaining clean stock is the only method by which its eradication may be guaranteed. If the value of the strain of animals justifies the effort, the entire colony may be divided into small, isolated groups. Then, those in which the disease continues may be destroyed, and other groups tested for fecal carriers[4] Fecal carriers occur most commonly among rats and mice, but fecal examination of guinea pigs and rabbits is not reliable. Skin testing has been used instead.[3]

Whether or not it may be possible to eradicate *Salmonella* infections in stocks of valuable animals, by use of one of the new antibiotic substances, remains to be determined.*

2. Pasteurellosis. Two forms of pasteurellosis occur in laboratory animals, one caused by infection with *Pasteurella aviseptica,* and the other by *P. pseudotuberculosis.* A striking feature of *P. aviseptica* infection in some species is its rapidly fatal course. In such animals, wide-spread capillary hemorrhages and congestion of viscera are conspicuous features. Hence, its designation, "hemorrhagic septicemia." *P. pseudotuberculosis* infection, usually, is a more chronic disease accompanied by the formation of circumscribed tubercle-like nodules in lymph nodes, lungs, and spleen. Because of this lesion, the disease has been called "pseudotuberculosis," a rather unfortunate term, since other organisms cause the development of similar abscesses.

*In reference to control of *Salmonella* infectious disease, Dr P A Mattis has called my attention to the use of sulfasuxidine and sulfthalidine in intestinal infections Available publications do not mention tests against paratyphoid organisms, but these drugs may be fed to rats and mice at 2–4% levels in the diet, without evidence of injury[9a, 9b]

[9a] **Thorp, T. S., V. J. Pisciotta, & Cora B. Grundy.** J. Am Vet. Med. Assoc. **104:** 274–278. 1944.

[9b] **Mattis, P. A., W. M. Benson, & H. S. Koelle.** J. Pharm. and Exp. Thera. **81:** 116–132. 1944.

A. Hemorrhagic septicemia. *P. aviseptica,* the cause of "hemorrhagic septicemia" is a parasite of many species of wild and domesticated animals. Some texts list almost as many specific names for this organism as it has hosts. Thus, its nomenclature has been confused, but there seems to be only a single species. Among laboratory animals, rabbits, mice, and canaries are highly susceptible. Guinea pigs and rats are more resistant, but endemic and epidemic infections have occurred in both species, chronic infection being limited to the nasal passage.

Among laboratory animals, *P. aviseptica* infections probably are best known in rabbits in which a subacute or chronic respiratory disease, known as "snuffles," is the common manifestation. Less frequently, colonies of guinea pigs and rats, also, may carry chronic infections, the organisms being limited to the nasal passages. In mice and canaries, the disease occurs only as epidemics.

P. aviseptica is spread mainly by contact with carriers which discharge the organism in nasal and ocular secretions, but the organisms can be airborne or be carried in food. Crowding and unhygienic conditions obviously promote infections, and rapid spread seems to enhance the virulence of the organism.

In the more susceptible species, *P. aviseptica* infections may cause death before any but the earlier signs of sickness are noted. At *post mortem* examination, all viscera are congested and numerous capillary hemorrhages are seen on serous surfaces.

Less rapidly fatal infections are more common in rabbits, guinea pigs and rats. The acute forms of "snuffles" in the rabbit is characterized by purulent rhinitis and conjuctivitis. The nose and eyes are encrusted with exudate and breathing is labored. The animals are weak, appetite is lost, and emaciation is rapid. At autopsy, one finds purulent bronchitis, broncho-pneumonia, hemorrhagic pleural exudate and sometimes pericarditis. In rabbits which pass into the chronic form of the disease, the rhinitis continues. The animals are emaciated, lack appetite, and cough and sneeze. Autopsy of cases of this form of pasteurellosis reveals more or less extensive pneumonia and abscesses in the subcutaneous tissues, or, if chronic, "snuffles" may be a prolonged process with little or no pulmonary involvement, the infection being confined to the nasal passages.

Guinea pigs react much as do rabbits to *P. aviseptica,* some dying of the acute septicemic disease, others showing respiratory infection. In severe respiratory infections, one finds pneumonia, and hemorrhagic

and fibrinous pleuritis, and pericarditis; but, in this species, the infection also spreads to the peritoneal cavity and uterus, and causes enlargement of the spleen.

P. aviseptica is a small gram-negative rod. It grows readily on ordinary media at a wide range of temperature. Diagnosis of infections is not difficult, especially during epidemics, when highly virulent strains of the organism are most often encountered. Subcutaneous injection of a small amount of the heart's blood of a suspected animal into mice usually causes death in less than 24 hours. Blood smears from the inoculated animal contain the organism in great numbers. When colored by Wright's or Giemsa's stain or by good quality methylene blue, the organisms appear darker at each end than in the central part. That is, they show bipolar staining. Isolation of the organism from chronic cases is not so certain, either by culture or animal inoculation, because carrier strains may be relatively avirulent.

Control of *P. aviseptica* infections is best accomplished by selecting stocks of rabbits and guinea pigs that are immune to "snuffles." Any additions to the colonies naturally should be quarantined and carefully examined for signs of the disease. Meyer[3] has outlined a routine that has given satisfactory control.

B. Pseudotuberculosis. *P. pseudotuberculosis* is a parasite of guinea pigs and sometimes of rats, in which it may cause epidemic disease. Infections of other animals are sporadic and usually inapparent.

The natural route of infection with *P. pseudotuberculosis* is the intestinal tract. Affected guinea pigs lose weight, develop diarrhea and die in three or four weeks. Autopsy reveals greatly enlarged lymph nodes in the peritoneal cavity, and in the retro-peritoneal and inguinal regions.

These nodes contain large and small abscesses filled by thick pus. Lymphatic tissue of the gut wall likewise contains many abscesses, and the liver and spleen are enlarged and studded with pale grey nodules which elevate the capsule. Sometimes also the abscesses are found in the lungs. Sporadic cases in other species usually are found by accident when the animal is examined after death. In such cases, the lesion usually consists of single nodules in the lungs or abscessed lymph nodes in the peritoneal cavity.

In the typical lesion, this organism is commonly found within cells. It may be identified tentatively in smears from the abscess by its characteristic bipolar staining, and it grows readily on simple media at a wide range of temperature.

So little is known of the epidemiology of this infection that recommendations for its control cannot be given. Obviously, its discovery, however, in a colony of guinea pigs would warrant their sacrifice and replacement by clean stock.

3. **Pseudotuberculosis of Mice.** *Corynebacterium kutscheri* (*murinum*). This disease of mice is infrequent and chronic, characterized by tubercle-like lesions in the lungs accompanied by pleural effusion. Similar tubercle-like lesions may be found in the liver, lymph nodes of the cervical region, the mediastinum, the mesentery, and isolated nodules in the spleen and kidneys. The organism can be recovered from any of the lesions. The chief interest in this disease is its similarity to Salmonella infections.

4. **Infectious Catarrh.** Chronic respiratory disease of rats is relatively common in many colonies. The disease involves the nasal passages, the middle ear, and the lungs. A high percentage of rats over 600 days of age show some degree of pneumonia. A similar disease occurs in mice also, and, in some instances, has been rapidly progressive.[10] The first sign of this condition in rats may be disturbances of equilibrium. More often, animals first appear unthrifty, after which breathing becomes increasingly labored. In one epidemic among mice, the first signs of illness was described as "chattering,"[10] followed by the development of rhinitis and pneumonia.

Many species of bacteria have been isolated from the nose, ears, and lungs of rats and mice affected by this disease, but none of them appears to be the primary cause. The recent work of Nelson,[10, 11] has shown that these organisms are contaminants and that the disease is caused by minute gram-negative coccobacilliform bodies which can be cultivated only in tissue culture. These reproduce the disease upon intranasal inoculation.

This infection is spread by contact early in the life of the animal. At least two methods to control it have been devised. One is by selective breeding for resistant strains.[12] The other is by using healthy foster-mothers for new-born young of diseased parents.[13]

5. **Tuberculosis.** Sporadic cases of tuberculosis may occur in colonies of guinea pigs and rabbits, but, under usual circumstances, these animals have little if any chance of infection. In monkeys, however, tuberculosis is a more important problem.

---

[10] Nelson, J. B. J. Exp. Med. 65: 833–841; 843–849; 851–860. 1937.
[11] Nelson, J. B. J. Exp. Med. 72: 645–654. 1940.
[12] King, H. D. Anat. Rec. 74: 215–222. 1939.
[13] Nelson, J. B., & J. W. Gowen. J. Exp. Med. 54: 629–636. 1931.

All species of monkeys and apes that are commonly used in laboratory work, as well as many that are not, are highly susceptible to tuberculosis. Moreover, conditions under which they are imported insure infection of a significant number.

Incidence of tuberculosis in newly imported monkeys varies from 10 to 25 per cent when large numbers are received in one shipment, and infection spreads readily, if the animals are confined in large groups in small cages. The route of infection is either respiratory or intestinal. Autopsy of cases that are allowed to run their course often fails to determine the portal of entry.

The average survival time of monkeys that have been infected naturally probably is about six months. The disease is not distinctive. The only reliable method of diagnosis during life is either one of two forms of the tuberculin tests.

The simplest tuberculin test for monkeys is the ophthalmic or palpebral test as developed by Schroeder.[14] This consists in a single injection of 1 milligram of Old Tuberculin or of the Purified Protein Derivative into the subcutaneous tissue of the upper eye-lid. Positive reactions appear between 48 and 72 hours. If the animal has tuberculosis, the eye-lid swells conspicuously. Otherwi , there is no response. This test is approximately 85 per cent accurate as determined by autopsy.[15]

The more complex tuberculin test, as devised by White and Fox,[16] is an adaptation of the original tuberculin test for cattle. It requires that the rectal temperature of each animal be taken at 3:00 P. M. for four days before subcutaneous injections of Old Tuberculin. On the morning of the fifth day, animals are injected and rectal temperatures taken at four-hour intervals for 48 hours. Abnormal temperature reactions indicate tuberculosis, but even a considerable experience in interpreting these reactions does not sustain one's self-assurance quite so well as does a monkey with a swollen eyelid.

The lesions of tuberculosis in monkeys are found mainly in the lungs, spleen and lymph nodes. In the lungs, one finds large and small areas of caseation necrosis, or actual pus formation. Sometimes, a whole lung is consolidated, and, on section, made up of alternate areas of caseous necrosis and darker, more vascular zones of recent involvement. Tracheo-bronchial lymph nodes are usually very large and necrotic except for a narrow rim of peripheral tissue. When infection enters by way

[14] Schroeder, C. R. Zoologica 23: 397–400. 1938.
[15] Kennard, M. A., & M. D. Willner. Yale J. Biol. & Med. 13: 795–812. 1941.
[16] White, C. Y. & H. Fox. Arch Int. Med. 4: 517–528. 1909.

of the intestinal tract, the mesenteric nodes show the major degree of enlargement. Tubercles in the spleen may vary from 0.1 to 1.0 centimeter, while in the liver the smaller size is the rule. In any of these lesions, tubercle bacilli usually are abundant and may be found in a smear without prolonged search.

Control of tuberculosis in monkeys depends upon elimination of all reactors and exclusion of all sources of infection. By testing a colony at six-month intervals for 18 months or two years, and, thereafter, at yearly intervals, the disease may be eliminated, if, of course, personnel in charge of the animals is free of infection, and all new additions to the colony are free of disease.

6. **Miscellaneous Bacterial Disease.** Laboratory animals are subject to many other bacterial infections, some of which may occasionally cause epidemics. Rarely are these forms parasites of more than one species of animal. The more important of these will be discussed here.

"Mouse Septicemia,"[4] designates a relatively infrequent, slowly progressive disease caused by *Erysipelothrix muriseptica*. The chief manifestations of this infection are conjunctivitis, pneumonia, edema of abdominal tissues, enlargement of the spleen, and circumscribed small pale grey areas of necrosis in the liver. This organism of mice cannot be distinguished from *E. rhusiopathiae,* which causes swine erysipelas or "diamond-skin disease." Since mice are commonly employed as test animals for the diagnosis of swine erysipelas, there is a possibility of confusion. *Erysipelothrix* may be isolated by inoculating mice introperitoneally with emulsions of suspected lesions. If the material be infected, death follows in one to four days, the blood and spleen containing large numbers of the organism. It is a gram-positive, small slender rod, straight or slightly curved.

PYOGENIC INFECTIONS

Mice, rats and guinea pigs often develop abscesses of the subcutaneous tissues of the head, neck and other parts of the body. Often, the regional lymph nodes are also diseased. At times, the condition becomes frequent enough to be called epidemic especially among guinea pigs.[17] Streptococci, staphylococci, and other organisms have been isolated. Apparently, these bacteria gain entrance to the tissues through small surface wounds. The simplest method of dealing with this condition is to destroy affected animals and reduce contamination of the cages and animal houses.

[17] Megrail, E. & E. H. Hoyt. J. Bact. 17: 54. 1929.

In conjunction with these incidental diseases of small rodents, it may be well to consider briefly infection by *Streptobacillus moniliformis*. Strains of this organism are commonly found in the nasopharynx of rats. It is nonpathogenic for this species, but, in mice, it produces a highly fatal generalized infection. Essentially, this disease is characterized by multiple abscess formation in the joints and internal organs. At times, however, the infection may be a rapidly fatal septicemia. Other strains of this organism also cause abscesses in guinea pigs. The relationship of *S. moniliformis* to disease of rats, mice and guinea pigs is complicated by its association with organisms of the pleuro-pneumonia group. The pleuro-pneumonia organisms have been associated with polyarthritis of rats, with rolling disease of mice, and, in combination with *S. moniliformis*, as a cause of pyogenic infection of guinea pigs. Sabin's[18] monograph should be consulted for a complete discussion of these organisms.

PNEUMONIA

Epidemics of pneumonia caused by pneumococci and Friedlander's bacillus have been described in colonies of both guinea pigs and mice. In both species, mortality was high. However, such outbreaks of disease apparently are rare and occasioned by unusual circumstances. These animals ordinarily are not carriers of either type of organism.

In other instances, *Brucella bronchiseptica* has been isolated from epidemic pneumonia of mice and guinea pigs, and from pulmonary disease of rats. This organism has long been known as a contaminant of "snuffles" of rabbits, and distemper of dogs, and there is considerable doubt that it is naturally pathogenic for other animals.

BARTONELLOSIS

(Ref. 4, 5, 19, 20)

Organisms of the genus *Bartonella* are usually grouped with the bacteria, although their position is uncertain. Infections by these organisms occur widely in rats, mice, guinea pigs, and dogs, but become apparent only after splenectomy or some other serious disturbance of the immune mechanism. Then they apparently cause severe and sometimes fatal anemia. *Bartonella* is transmitted by blood-sucking arthropods, usually lice. Colonies free of lice are free of bartonellosis. Somewhat similar to *Bartonella*, are organisms assigned to the genera

[18] **Sabin, A. B.** Bact. Rev. **5**: 1–16. 1941.
[19] **Weinman, D., & H. Pinkerton.** Ann. Trop. Med. and Parasitol. **32**: 215–217. 1938.
[20] **Groot, H.** Proc. Soc. Exp. Biol. and Med. **52**: 279. 1942.

Eperythrozoon and *Grahamella*. These are parasites of mice. Apparently, they are not pathogenic.

FUNGUS INFECTIONS

(Ref. 4, 5)

Fungus infections of laboratory animals usually are limited to the skin. They occur most often in dogs and cats—street animals that have many opportunities of contact. However, infections are well known in kennels, and spread easily. All of the rodents are susceptible and epidemics may occur among them. The common dermatotropic fungi are assigned to the genera *Microsporon* and *Trichophyton*. Lesions consist of small, circular, denuded, thickened, scaly patches distributed over much of the body, or larger, irregular, denuded areas that are reddened, moist and scaly. Spontaneous recovery sometimes occurs. It can be hastened by application of tincture of iodine or five per cent alcoholic solution of salicylic acid.

SPIROCHETAL INFECTIONS

(Ref. 3, 5)

Infections of laboratory animals by spirochetes are of little practical significance; but, perhaps, it may be well to note that natural infections occur both in dogs and rabbits. The incidence of *Leptospira icterohemorrhagiae* and *L. canicola*, organisms which cause jaundice and renal disease in dogs and Weil's disease in man, is approximately 20 per cent in this country.[21] This is something of a hazard to laboratory workers, but probably not a very great one. Spirochetosis of rabbits is a benign disease caused by *Treponema cuniculi* transmitted by contact, coitus, or inoculation.

VIRUS DISEASE

One is tempted to conclude that the virus infections of laboratory animals were designed more as a source of confusion than of morbidity. Under ordinary circumstances, few of these agents cause significant disease, but, when their hosts are inoculated with other agents or even stimulated by inert substances, the infection becomes apparent.

Probably, because of the great number of mice that are being used for experiment, a greater number of virus infections have been recognized in this than in any other animal.[22] Many of these are thoroughly discussed in Dingle's[4] review, but others have been re-

---

[21] Greene, M. R. Am. J. Hyg. 34: 87–90. 1941.
[22] Thompson, J. Arch. Path. 21: 531–540. 1936.

ported more recently. I shall give only brief mention to these infections, listing them according to the chief regions of localization. One will, of course, realize that many viruses invade the body widely, although symptoms of infection may be related to one or another group of organs or tissues.

Neurotropic Viruses: (1) Mice. Encephalomyelitis, Theiler,[23] flaccid paralysis of the hind legs, but not of the tail; incidence low, young animals chiefly. Meningoencephalomyelitis,[24] paralysis of hind legs; incidence unknown. Lymphocytic choriomeningitis:[25] Symptoms indefinite, young affected, morbidity and mortality low. Infected mice said to be more susceptible to intercurrent disease and lymphomatosis than other strains.[26] (2) Monkeys. Lymphocytic choriomeningitis[27] and "B" virus;[28] incidence unknown, inapparent infections until stimulated by other agents. (3) Guinea pigs. Ascending paralysis;[29] incidence unknown. May be caused by virus of salivary gland disease.[30]

Pneumotropic Viruses: (1) Mice. Virus pneumonia:[31, 32, 33, 34, 35] Many types, inapparent infections until stimulated by repeated passages. (2) Guinea pigs. Virus pneumonia:[36] Highly fatal disease, spread by contact; incidence unknown. (3) Hamsters:[37] (*Cricetus aruatus*) Virus pneumonia: Latent infectious until stimulated by repeated passage, incidence unknown. (4) Cats [38, 39] Virus pneumonia: Highly infectious in young animals, mortality low.

Viscerotropic viruses: (1) Guinea pigs. Entero-hepatitis:[40] Acute disease, diarrhea, wasting; incidence unknown. Pneumonia-hepatitis:[41] A slowly progressive disease; incidence unknown. (2) Mice. *Infectious Ectromelia:*[4] Acute disease, often passing into chronic form. Acute state; polyserositis, and acute hepatitis: Chronic stage; edema, vesiculation and ulceration of skin of feet, and amputation of feet; incidence unknown, probably widespread, but usually inapparent.

---

[23] **Theiler, M.** J. Exp. Med. **65**: 705–719. 1937.
[24] **Gahagan, L., & L. D. Stevenson.** J. Inf. Dis. **69**: 232–237. 1941.
[25] **Traub, E.** J. Exp. Med. **63**: 533–546. 1936.
[26] **Traub, E.** Zentralbl Bakt. I. Abt. Orig. **147**: 16–25. 1941.
[27] **Armstrong, C., & R. D. Lillie.** U. S. Pub. Health Rep. **49**: 1019–1027. 1934.
[28] **Sabin, A. B., & A. M. Wright.** J. Exp Med. **59**: 115–136. 1934
[29] **Romer, P. H.** Deutsche Med. Wohnschr. **57**: 2685–2689 1910.
[30] **Cole, R., & A. G. Kuttner.** J. Exp. Med. **44**: 855–873. 1926.
[31] **Horsfall, F. L., & R. G. Hahn.** J. Exp. Med. **71**: 391–408. 1940.
[32] **Herzberg, K., & W. Gross.** Zentralbl. Bakt. I. Abt. Orig. **146**: 129–239. 1940.
[33] **Freeman, G.** Proc. Soc. Exp. Biol. and Med. **48**: 568–610. 1941.
[34] **Nigg, C.** Science **95**: 49–50. 1942.
[35] **Karr, H. V.** J. Inf. Dis. **72**: 108–116. 1943.
[36] **Ten Broeck, C., & J. B. Nelson.** Proc. Soc. Exper. Biol. and Med. **39**: 572–575 1938.
[37] **Pearson, H. E., & M. D. Eaton.** Proc. Soc. Exper. Biol. and Med. **45**: 677–679. 1940.
[38] **Baker, J. A.** Science **96**: 475. 1942.
[39] **Thomas, L. & E. M. Kolb.** Proc. Soc. Exper. Biol. and Med. **54**: 172–174 1943.
[40] **Marcos, E.** J. Bact. **25**: 239–240. 1933.
[41] **Pappenheimer, A. M. & C. A. Slanetz.** J. Exper. Med. **76**: 299–306. 1942.

(3) Dogs. Distemper:[5] Acute generalized disease. Enteritis, or encephalitis may predominate; pneumonia secondary; world-wide distribution, highly contagious, often fatal; control by vaccination. (4) Cats. Panleukopenia:[42,43] Acute generalized disease, world wide distribution, highly contagious, often fatal; control by vaccination.

Epidermotropic Viruses: (1) Rabbits. Pox:[44,45] An acute disease, highly contagious, relatively uncommon, lymphadenitis, papules on skin and mucus membranes, keratitis, ophthalmia, orchitis; recovery prolonged, often fatal to young animals. "Virus III":[46] An inapparent wide-spread infection, produces skin lesions when transferred in series. Oral papillomatosis:[47] Common benign disease of buccal mucosa, transmitted by contact. (2) Dogs. Oral papillomatosis:[5] Highly contagious disease, spread by contact, causes serious inconvenience to the animal. (3) Canaries. Pox:[48,49,50] Highly fatal contagious disease, buccal muccosa, and viscera.

Salivary Gland Disease: An inapparent infection of guinea pigs, mice, rats, Chinese hamsters, and monkeys; inclusion bodies in hypertrophied cells of the ducts, of the salivary gland and sometimes kidneys.[51,52]

ANIMAL PARASITES

Coccidia, lice, and mites are the more important animal parasites of laboratory animals. One should be aware, however, that many other animal organisms, both protozoa and metazoa, may be found in the intestinal canal and solid organs of all species. Rats, mice and rabbits are readily infected by the larvae of dog and cat tapeworms. These parasites are introduced into animal colonies in bedding or food contaminated by dog and cat feces. Rats and mice, also, are readily infected by nematodes and cestodes, native to their wild relatives.[53] Street dogs and cats are commonly parasitized by several species of roundworms and cestodes. Monkeys, often, are infected by *Strogyloides, Oesophagostomum* and *Endamoeba histolytica*. With the exception of Coccidia and certain other protozoa, however, these parasites

---

[42] **Hammon, W. D., & J. F. Enders.** J. Exper. Med. **69:** 327–352. 1939.
[43] **Kikuth, W., R. Goennert, & M. Schweickert.** Zentralbl. Bakt. I. Abt. Orig. **146:** 1–17. 1940.
[44] **Green, R. G.** J. Exper. Med. **60:** 427–440; 441–455. 1934.
[45] **Pearce, L., P. D. Rosahn, & C. K. Hu.** J. Exper. Med. **63:** 241–258. 1936.
[46] **Rivers, T. M., & W. S. Tillett.** J. Exper. Med. **40:** 281–287. 1924.
[47] **Parsons, R. J., & J. G. Kidd.** J. Exper. Med. **77:** 232–250. 1943.
[48] **Shattock, S. G.** Lancet **1:** 1055. 1898.
[49] **Kikuth, W., & H. Gollub.** Zentralbl. Bakt. I. Abt. Orig. **125:** 313–320. 1932.
[50] **Burnet, F. M.** J. Path. and Bact. **37:** 107–122. 1933.
[51] **Kuttner, A. G., & S. Wang.** J. Exper. Med. **60:** 773–791. 1934.
[52] **Thompson, J.** J. Inf. Dis. **58:** 59–63. 1936.
[53] **Hall, M. C.** Proc. U. S. Nat. Mus. **50:** 1–258. 1916.

either may be eliminated easily by treatment or their invasion of the colony, prevented completely by application of simple hygienic measures.

PROTOZOAN INFECTIONS

Coccidiosis: [54] As a rule, coccidiosis is a disease of young, growing animals. Among the laboratory animals, its severest forms occur most commonly in rabbits; but, under unusual circumstances, other animals may be seriously affected.

Coccidia present striking examples of host specificity and, sometimes, of organ specificity. Rarely are species of these protozoa interchangeable between different animal species. Thus, control is not so complex as it is with some of the pathogenic bacteria; but, even so, factors determining the spread of the infection and its severity remain unchanged. That is: the parasite is maintained by carriers, it is spread by contact and, on the whole, severity of disease is influenced both by the size of the infecting dose and the susceptibility of the animal.

Coccidiosis of rabbits involves both the liver and the intestines. Either or both forms of the disease may occur in the same host. One is no protection against the other, and both types may vary from rapidly fatal infections to mild inapparent disease. In either hepatic or intestinal coccidiosis, severe disease is accompanied by diarrhea and wasting, and subacute or chronic cases of hepatic coccidiosis are accompanied by marked abdominal enlargement.

Post mortem examination of animals dying of hepatic coccidiosis finds the liver, gall bladder and bile ducts hypertrophied; and, if the disease has been subacute or chronic, many white or yellow nodules, up to two centimeters in diameter, are scattered over the liver. Depending upon their age, the contents of these lesions vary from thin whitish fluid to thick caseous material, which later may become calcified and mistaken for tubercles. In animals dying of intestinal coccidiosis, the small intestine is dilated and pale, and the mucosa may be thickened.

Hepatic coccidiosis of rabbits is caused by *Eimeria stiedae;* intestional coccidiosis by *E. perforans, E. magna* and possibly other species. Infections may be recognized before death by finding the characteristic oocysts in the stools. Identification of species requires careful measurement. At autopsy, oocysts may be found in large numbers in the

[54] **Becker, E. R.** Coccidia and Coccidiosis of Domesticated Game and Laboratory Animals and of Man. Collegiate Press, Inc., Ames, Iowa. 1934.

early nodules of the liver, but are more difficult to identify from the intestinal epithelium.

Control of coccidiosis is difficult, because the parasite is so widely distributed and is transmitted from mother to offspring so readily. Modern hutches which prevent fecal contamination of food and the cage, or at least reduce it to a minimum, do much to prevent losses from this disease.

Dogs, cats, rats, mice, and guinea pigs, also, are parasitized by one or more species of coccidia; but, on the whole, these animals are not often seriously affected by the organisms. If, however, they are fed large doses of oocysts, dogs, rats and guinea pigs may develop severe enteritis.

Other Protozoan Infections: In addition to coccidia, many other species of protozoa occur regularly in the intestines of rodents, rabbits and monkeys. With the exception of *Endamoeba histolytica,* a common parasite of many species of monkeys and rarely of dogs, none of these organisms cause disease, and the importance of *E. histolytica* may be questioned. I have seen a large number of monkeys that carried *E. histolytica* without evidence of disease either before or after death, and have seen it cause disease only in animals of the genera *Ateles* and *Lagothrix.*

Protozoan parasites of the blood and tissues also may occur among laboratory animals. Sometimes, they may cause disease, but possibly more often, they cause confusion. One of the most common is *Toxoplasma gondii*[5, 55, 56] an organism of undetermined position, but with a wide range of hosts. It has been found in mice, rats, guinea pigs, rabbits, dogs, cats, sheep, man and many species of birds. In some instances—rabbits, and man—it has been associated with encephalitis, but again it has been found only when brain tissue was inoculated into other animals. The infection in cats and dogs involves tissues other than the brain.

Somewhat similar to *Toxoplasma* are the Sarcosporidia. These are parasites of muscles of a variety of animals. In the muscles, one may find tiny white tubular structures visible to the unaided eye or, again, the parasite is found by chance in section of muscle as a thin-walled tube containing bodies three to seven microns in length. Apparently, these organisms do not cause active disease.

It also may be well to mention *Babesia canis,* a parasite of dogs throughout the warmer parts of the world. This organism is trans-

---

[55] Sabin, A. B., & P. K. Olitsky. Science 85: 336–338. 1937.
[56] Wolfe, A., D. Cowen, & B. H. Paige. Amer. J. Path. 15: 657–694. 1939.

mitted by several species of ticks, and animals that are exposed early in life usually suffer little, but in non-immune dogs, there is acute disease, with high fever, anemia, jaundice and hemoglobinuria. Thus, the disease simulates leptospirosis. *B. canis* is a parasite of erythrocytes and is easily transmitted by transfusion.

METAZOAN PARASITES

The cestodes and nematodes that parasitize laboratory animals are not particularly significant. However, it may be well to recall that the hookworms of dogs are widely distributed and, since they are blood-sucking organisms, may interfere with experiments. Dogs may be infected by any one of three species of hookworms: *Ancylostoma caninum, A. braziliense,* and *Uncinaria stenocephala.* Infections are recognized by finding characteristic ova in the stool. Stool examination is greatly facilitated by concentration or flotation methods.[57]

Strongyloides stercoralis, another nematode parasite of the intestine, may occur in dogs, cats and monkeys and be associated with diarrhea. The larvae of this species are found in the feces. Monkeys are also parasitized by species of *Oesophagostomum,* which produces inflammatory nodules in the wall of the colon and may cause death by perforation of one of these lesions. This parasite is rapidly lost if reinfection be prevented by reducing fecal contamination of the food.

Acarasis: All species of laboratory animals may be infected by mites, some of which merely hide in the cages and take blood at night, but others burrow into the skin of the head, ears and body causing considerable damage to their hosts. *Sarcoptes scabiei* causes one form of mange in dogs, *Demodex canis* another, and species of *Otodectes* infect the ears of dogs, cats and occasionally rabbits; but *Psoroptes cuniculi* is the more common cause of ear mange in rabbits. However, there is little profit simply in listing the names of the various species of mites and lice that may be found on laboratory animals. Lesions produced by the mange mites are all more or less similar. The skin is bare, thickened and covered by crusts. The parasites may be found by scraping the lesions deeply, emulsifying the crusts in water or sodium hydroxide solution, and examining them under the microscope, or the larger mites may be seen with the hand lens or unaided

[57] Coffin, D. L. Manual of Veterinary Clinical Pathology. Comstock Publishing Co., Inc., Ithaca, N. Y. In Press.

eye. Identifications may be made by reference to any good text, such as Mönnig's "Veterinary Helminthology and Entomology."[58]

Control of mites that do not invade the skin may be accomplished by spraying the cages with light volatile oils, or by sterilizing cages with steam or dry heat. Ridding animals of mange and follicle mites is considerably more of a problem and requires careful treatment of all affected animals at intervals of 7 to 10 days, as well as disinfection of cages, and animal rooms. Mönnig lists formulae that have proven useful in treatment.

In attempting this review, a major problem has been the choice of material to be included. An all-inclusive survey obviously would be impossible. A second question has been the space allowance for any one disease. An adequate consideration of the infectious diseases of laboratory animals, with illustrations of morbid processes and directions for diagnosis and control, would occupy a volume of several hundred pages. Much of this would merely be repetition of readily available information. Possibly, such a volume might be valuable to some, but the effort would be difficult to justify. The present survey can be regarded only as an introduction.

DISCUSSION OF THE PAPER

Dr. C. A. Slanetz (*College of Physicians and Surgeons, Columbia University, New York, N. Y.*):

Dr. Ratcliffe has reviewed the important infectious diseases of laboratory animals so adequately that I can add little except some of our experiences with disease control of small animal colonies at the College of Physicians and Surgeons Columbia-Presbyterian Medical Center. Our observations cover a period of 14 years. Although, for the most part, they fit the picture as presented by Dr. Ratcliffe, we have made, I believe, some observations which are different and which may be of interest to individuals concerned with the origin and care of experimental animals.

As indicated by Dr. Ratcliffe in his report, we also have found *Salmonella* organisms to be the most important cause of disease among mice, rats and guinea pigs. The success we have had in maintaining breeding colonies of rats, guinea pigs and mice practically free of paratyphoid infections, for the past ten years, has been very helpful in our work at the college. Our program of disease control includes testing of breeders at monthly intervals and, at times, more frequently. Fecal specimens are taken and streaked on brilliant green agar, containing a dye dilution of 1:260,000. No difficulty has been encountered in keeping our rat colony entirely free of infection.

With mice, the problem has been less simple. Occasionally, it has been necessary to culture the breeders more frequently than once a month to make sure that no paratyphoid carriers remain in the colony. *S. enteritidis* and *S. typhimurium* are the two species which we have found in most instances.

Our experience in large scale testing includes a commercial mouse colony which supplies a part of our needs. The results have been encouraging. At present, the

[58] Mönnig, H. O. Veterinary Helminthology and Entomology. 2nd. ed. Williams and Wilkins Co., Baltimore. 1941.

incidence of *Salmonella* infection in that colony is a fraction of one per cent, judging from tests on groups of 200 to 400 mice, compared to an average of 15 per cent before testing, was instigated five years ago. Stock mice from several commercial sources have shown an incidence of 5 to 30 per cent paratyphoid carriers.

So far, our guinea pig colony has not given us any difficulty with paratyphoid, even during the breeding period when others report *Salmonella* infections to be prevalent in pregnant females and in females at parturition.

We have not encountered paratyphoid in our rabbit breeding stock.

Control of rabbit coccidiosis has been difficult. By frequent examination of fecal specimens (flotation method) of breeding stock, selecting negative does, and the use of screen floors in cages, we had been able to reduce the incidence of liver coccidiosis in young stock from 90 per cent to approximately 5 per cent. The cost of such a program is considerable and is not practical on a large scale.

When an epidemic of snuffles occurs among our rabbits, the survivors usually prove resistant to snuffles in subsequent epidemics.

Tuberculosis in monkeys is no longer a problem with us. By housing healthy stock in individual metal cages with screen floors, solid sides, and proper care of food, we have been able to maintain a colony of 40–60 monkeys over a period of 8 years with only three deaths from tuberculosis. When we formerly housed monkeys in groups of 6–10, our losses from tuberculosis were high at the end of 6–8 months.

Dog distemper is still a disturbing factor in our dog experimental program. Our dogs are purchased from a dealer in Pennsylvania and are probably exposed to distemper at the kennels there or in contact with distemper cases in the shipping crates. By injecting small dogs and puppies with suitable doses of anti-canine distemper serum followed by the regular prophylactic vaccination we have been able to develop immunity in such animals. Large breeds of dogs, shipped to us from Pennsylvania, have not always responded satisfactorily to such treatment, possibly because of inadequate doses of serum or other factors.

Dr. C. R. Schroeder (*Lederle Laboratories, Pearl River, New York*):

The infectious and parasitic diseases of laboratory animals have been ably and amply covered by Dr. Ratcliffe.

Biological and pharmaceutical manufacturing plants are probably the largest users of small animals today for testing and experimental purposes. The infectious disease problem has great economic importance for this group. Hemolytic streptococcal adenitis and salmonellosis are the cause of greatest losses.

Specifically immune animals cannot be used because the common group of test animals must have equal susceptibility to all the genera of organisms, toxins and viruses to which they may be exposed.

The maintenance of large, confined colonies of laboratory test animals with rapid population turnover, free of interfering infectious disease, is a difficult management problem.

Control measures include:

1. Systematic testing to screen out carriers and prevent the introduction of infectious disease and parasites. Maintaining adequate quarantine and only admitting clean stock to breeding colonies and testing units.
2. Furnishing proper housing, including:
   a. adequate temperature and humidity control;
   b. avoiding over-crowding;
   c. furnishing suitable bedding;
   d. avoiding draughts and excessive noise.
3. Maintaining adequate diets.
4. Giving attention to good sanitary practices, including:
   a. proper disposition of sick and dead animals and waste;
   b. thorough cleaning and disinfection of buildings and cages.

Sanitation, including the use of mechanical cleaners, cleaning compounds and disinfectants, has not been discussed and it seems timely and in order to treat it here.

There are many types of cleaning units which will physically remove organic matter from any surface, regardless of its composition.

Three types of equipment are available:

1. Flash boiler steam generators:
   Steam pressure, in excess of 100 pounds, is generated to propel water and heat. An injector forces water and detergent in correct proportion into the steam line, so that an ample stream of water with detergent at 100° C. in excess of 100 pounds pressure is delivered to the nozzle at the end of a steam hose, to be directed at close proximity to the part to be cleaned.
   This type of unit may be stationary or mounted on a trailer to permit delivery to the job where plant steam is not available. A source of water, together with a 110 volt outlet, is needed. The boiler fuel is kerosene or the equivalent. (This equipment is distributed by the Homestead Valve Manufacturing Company, Coraopolis, Pennsylvania.)
2. A unit composed of pumps which build extremely high cold water pressures in excess of 400 pounds in 10 seconds (manufactured and distributed by the Drill Manufacturing Company, Circle Tower, Indianapolis, Indiana).
3. Units for steam cleaning, utilizing available plant steam of 75 to 150 pounds pressure. These units syphon a premixed solution from a portable reservoir and inject it into the steam line (manufactured and distributed by the Kerrick-Hydro Steam Kleaner, Clayton Manufacturing Company, Alhambra, California, and Oakite Products Inc., 46 Thames Street, New York City).

The following manufacturers may be consulted in selecting the proper detergent to be used in conjunction with the above equipment for efficient cleaning.

Dextrex Corporation, 13013 Hillview Avenue, Detroit, Michigan; Kemico Mfg. Company, Irvington, New Jersey; Oakite Products, Inc., 46 Thames Street, New York City; Turco Products, 48th and South Holsted, Chicago, Illinois; U. S. Sanitary Specialties Corporation, 437 South Western Avenue, Chicago, Illinois; Wyandotte Chemicals Corporation, Biddle Avenue, Wyandotte, Michigan.

Final disinfection can be accomplished by using any of the quarternary ammonium compounds like Roccal. Chlorine compounds such as H T H, perchloron, or chlorinated lime, may be used, but they are destructive to metal parts. Mercurial and coal tar preparations should be avoided. Usually, thorough cleaning with ample water and detergents precludes the necessity of a complex disinfection program.

INFLUENCE OF ENVIRONMENTAL TEMPERATURES ON WARM-BLOODED ANIMALS

By Clarence A. Mills

From the Department of Experimental Medicine, University of Cincinnati, Ohio.

Studies of recent years have gradually exposed the dominance which environmental temperatures exert over the existence of all warm-blooded animals. Ease of body-heat loss is now seen to play a major role in determining such vital factors as growth rate, speed of development, fertility, resistance to infection, dietary requirements for certain vitamins, and even mental ability. In order to understand just why and how this should be so, it is necessary to consider the animal body as a combustion machine.

Dynamics of warm-blooded existence. All body functions require energy for their performance, and the only source for such energy is the cellular combustion of foodstuffs—chiefly glucose. Unfortunately, various studies have shown homothermic animals and man to work at relatively low efficiency, being able to transform only 20–25% of their combustion energy over into work output. This is similar to the working efficiency of a good gasoline motor but far below the 37+% working efficiency now achieved by the Diesel engine.

This low efficiency of the animal body means that 75–80% of all combustion energy must be dissipated as waste heat; and, since heat accumulation (fever) seriously disturbs many important functions, it is necessary that this dissipation be readily accomplished. Hence it is that environmental temperatures—controlling, as they do, the ease or difficulty of body-heat loss—assume a truly dominant role in the existence of warm-blooded animals and man.

Variations in metabolic rate. Figure 1 shows variations in the basal metabolic rate of man which take place in southern Germany (at Heidelberg), as outdoor mean temperatures rise and fall through the seasons. Gelineo[1] has demonstrated a similar metabolic responsiveness in experimental animals under controlled conditions, as has also Lee.[2] These changes in basal or resting metabolism require 2 to 3 weeks for their appearance after environmetal temperatures have risen or fallen. The more immediate response to temperature change involves skeletal muscle tone with relaxation in heat and greater muscular activity or shivering in cold.

[1] **Gelineo, S.** "Influence du milieu thermique d'adaptation sur la thermogénèse des homeothermes." Ann. de Physiol. **10:** 1083 1934

[2] **Lee, Robert C.** Heat production of the rabbit at 28° C as affected by previous adaptation to temperatures between 10° and 31° C. J. Nutr. **23:** 83 1942

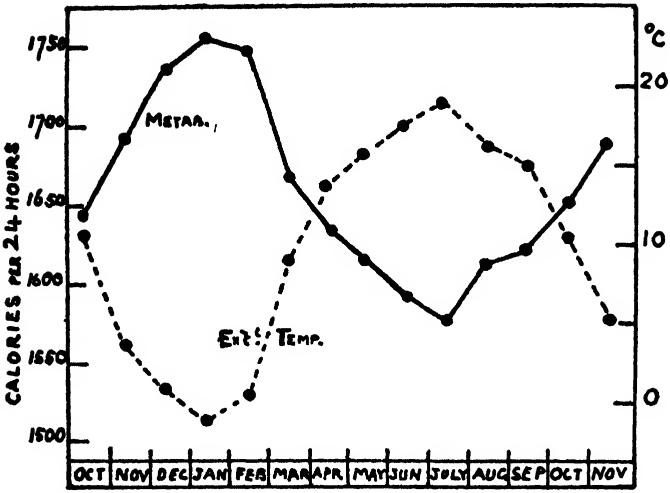


FIGURE 1 Mean monthly metabolism and mean monthly temperature Gessler (1925), observations on himself

The slower changes in basal metabolism should be kept constantly in mind. They bring profound alterations in nutritional requirements and the functions of many organs in animals grown or kept for experimental studies. Temperature control in animal quarters is now recognized as a necessary factor in careful research.

Growth and development. FIGURE 2 illustrates the typical growth retardation brought by tropical or summer heat, with mice receiving Purina chow *ad libitum* No difference is seen during the first week of heat exposure, but, beginning in the second week and continuing on indefinitely, there occurs a 40% growth deficit in the heat Only about half as much is eaten by the hot-room animals, so their metabolic efficiency is actually greater, giving more weight gain for each gram of food eaten. After 2–3 months, mice and rats in the heat are found to have tails considerably longer than those of their litter mates in the cold. Rabbits grow distinctly larger ears to facilitate heat loss when environmental temperatures are kept high. These developmental adaptations in heat-loss organs reflect quite clearly the metabolic handicap such warm-blooded animals face in hot surroundings. Within 2–3 weeks after heat-adapted animals are removed to cool surroundings, they resume a normal food consumption and growth rate.[8]

[8] C. A. Medical Climatology. C. C. Thomas, Springfield, Mass. 1939.

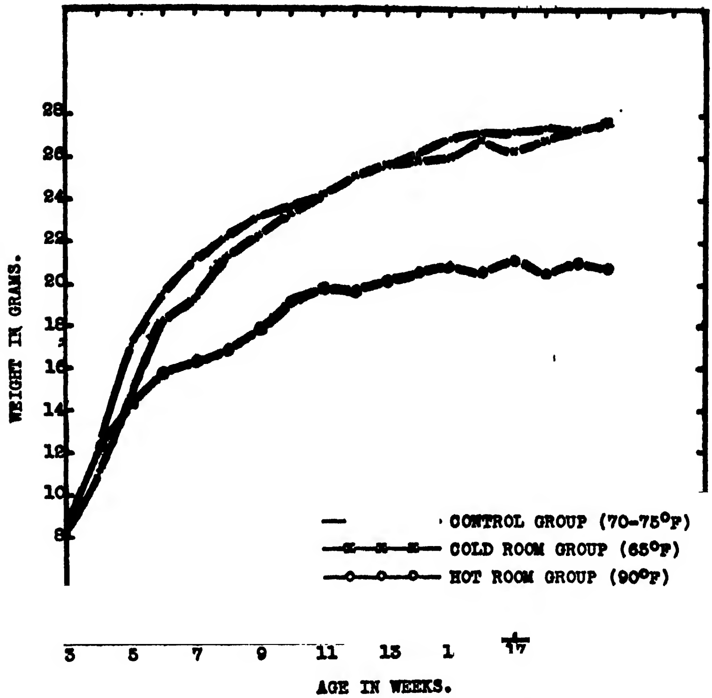


FIGURE 2. Growth of white mice at different temperatures

Onset of sexual functions and fertility. Mice kept at 90–91° F., from weaning age on, begin their sexual cycles significantly later than litter mates kept in cool quarters. Fertility in both males and females also comes on later in the heat and is, at all times, at a lower level of effectiveness. Mating takes place just as readily in hot as in cool environments, but conceptions are difficult to achieve in the heat, while almost every mating is effective in the cold. Litter size is significantly larger at 70° F. than at 90° F., with fewer still-births or deaths during the first week of life. Second-generation mice in the heat are almost completely infertile. Transfer from heat to cold usually results in a restoration of normal fertility within 3 weeks. Over-heating, with hyper-pyrexia for some hours, may result in permanent loss of fertility, even though the animal appears to be in perfect health otherwise.

Ability to store and mobilize glycogen. Liver glycogen is usually found to be less in heat-adapted animals, but more striking is their

inability to mobilize the stores they do have when exposed to body chilling. Those kept in cool quarters exhaust their liver glycogen almost completely in an effort to maintain a normal body temperature under chilling conditions, but those adapted to heat suffer a severe fall in rectal temperature and are prostrated by the cold, without having used more than a small fraction of their stored glycogen. This inability to meet chilling renders heat-adapted animals very susceptible to pneumonia when exposed to varying temperatures.

Mental Functions. While most of us are not often concerned with the thinking capacity of our laboratory animals, there are certain studies which deal directly with this phase of physiology.[4] It is thus important to know that environmental temperatures and ease of heat loss markedly affect mental functions. Rats adapted to different temperatures (55° F., 75° F. and 92° F.) for 3 months were tested for their ability to learn their way through a maze to food, after a 24 hours fast. Those kept at 55° F. required 12 trials and made few mistakes subsequently; those from the 75° F. quarters learned only after 28 trials and still made many mistakes; many of the rats adapted to 92° F. never did find the way to the food, while the successful ones found the food only after an average of 53 trials! Retested again, after a month's rest, those from the 55° F. room were found to have perfect retention of their previous learning, those from the 75° F. room had to relearn about half, while the previously successful ones from the heat showed no evidence whatever of their former experience. Thus, the ability to acquire and retain learning seems sharply dependent upon the ease of body-heat loss and the metabolic level allowed the animal. I might say that college students show similar heat effects!

ENVIRONMENTAL TEMPERATURES AND NEOPLASTIC TENDENCIES

Underfeeding, with the resultant malnutrition, has been shown to reduce the incidence of spontaneous tumors and leukemia in susceptible mice. Environmental warmth exerts a similar effect with mice on a Purina chow diet, perhaps because of a like lowering of food consumption and growth retardation. Both dba and C_3H strain mice show a significant drop in spontaneous breast tumor incidence, if raised at 90° F., as compared to 70° F.[5,6] The tumors occur at an earlier age, grow

[4] **Hellmer, Leo A.** The effect of temperature on the behavior of the white rat. Am. J. Psychology, **56**: 408–421. 1943.

[5] **Wallace, Edward W., Helene Wallace, & C. A. Mills.** Influence of environmental temperature upon the incidence and course of spontaneous tumors in C_3H mice. Cancer Research. **4**: 5 May, 1944

[6] **Fuller, R. H., Edna Brown, & C. A. Mills.** Environmental temperatures and spontaneous tumors in mice. Cancer Research. **1**: 130–133. 1941.

in size faster, and are more likely to be multiple in the mice adapted to stimulating coolness. Spayed females, while showing a marked reduction in breast-tumor incidence, still exhibit this same stimulating effect of a cool environment.

With chemically induced or transplanted tumors the story is quite different.[7] If methyl-cholanthrene be injected subcutaneously into C_3H mice, those previously adapted to 90° F. show a higher incidence of tumor development, earlier appearance and more rapid growth of the mass, and an earlier death ot the animal. An even more striking difference occurs with subcutaneous transplantation of tumor cells. In the heat, the incidence of "takes" is very high and growth of the mass rapid, while, in the cold, most implantations show only a slight period of growth before their regression and complete disappearance. With implantation into thigh muscles, however, there is noted no significant difference between hot- and cold-room tumor transplants.

These findings lead naturally to the conclusion that the level of tissue metabolism or richness of blood supply is probably an important factor in neoplasia. In hot surroundings, the rich blood supply to the skin for heat-loss purposes results in an active cutaneous metabolism and increased tendency to neoplasia. In cool environments and low skin temperatures, skin tumors are induced with greater difficulty, but spontaneous tumor incidence is higher in the deeper tissues and organs which have a more active metabolism in the cold. Whatever the mechanism by which the differences are produced, it is obvious that temperature control should be exercised in all tumor studies.

Resistance to infection. Animals adapted to 90° F. are much more likely to develop pneumonia, when chilled, than are those adapted to 70° F. Mice on Purina chow diet in the heat die significantly earlier after inoculation with a fixed number of Type I pneumococci than do those kept in the cold. With a less virulent organism (*Streptococcus*), dosages can be found which kill all heat-adapted mice, but fail to kill those from cool quarters. This lowered resistance in the heat is accompanied by only a slight reduction in ability to produce immune bodies or in the amount of immune serum required for passive protection.[8] It would therefore seem likely to be based upon differences in phagocytic activity in the heat and cold. Studies along this line will be mentioned in a later section dealing with nutritional requirements.

---

[7] **Wallace, E. W., Helene Wallace, & C. A. Mills.** Effect of climatic environment upon the genesis of subcutaneous tumors induced by methylcholanthrene and upon the growth of a transplantable sarcoma in C_3H mice. J. Nat. Cancer Inst. **3:** 99–110. 1942.

[8] **Mills, C. A., & L. E. Schmidt.** Environmental temperatures and resistance to infection. Amer. J. Tropical Medicine. **22:** 6. Nov. 1942.

Longevity. If the slower-growing mice in hot environments be carefully protected from chilling and infectious epidemics, they exhibit the same increased longevity that McKay obtained in his experiments on underfeeding. Voluntary underfeeding from loss of appetite and difficult heat loss in hot environments is probably quite analogous to the enforced malnutrition brought by limitation of food supply at ordinary temperatures.

Drug toxicity. Animals adapted to hot environments are hypersusceptible to the action of such drugs as insulin and thyroid extract.[9] The convulsive dose of insulin for mice is only 1/20–1/30, as much when they have been adapted to 90° F. as in cool surroundings (68° F.). With thyroid extract in rats, the difference is about 10-fold. Animals in the heat are, in general, much more susceptible to convulsant drugs such as strychnine and picrotoxin. Some of the sulfonamides have also been found more toxic in the heat. More studies are needed along this line.

Nutritional requirements. In an effort to counteract or prevent these unfavorable heat effects, we have made an extensive investigation of possible differences in nutritional requirements. First attention was centered upon the various B-vitamins, several of which are known to play essential roles in cellular combustion. It seemed possible that dietary enrichment with these combustion catalysts might make for greater metabolic efficiency, allowing the animal a higher level of existence in the continuing presence of difficulty in heat loss.

Briefly stated, it has been found that best growth is obtained in rats and mice kept at 90° F., when the dietary *thiamine* is twice as high (per gram of food) as is needed for optimal growth at 70° F.[10] For young rats, this means 0.8 mg. per kilogram of food at 70° F. and 1.6 mg. per kilogram at 90° F.; while, for mice, analogous requirements are about 2 mg. per kilogram and 4 mg. per kilogram. With advancing age, rats at 70° F. show no change in requirement, but those kept at 90° F. exhibit a progressive rise (to as much as 3 mg. per kilogram by 9 months of age). These differences in thiamine requirement, per gram of food in heat and cold, are not due to suppression of thiamine synthesis by intestinal bacteria in the heat, for they persist unchanged when 0.5% sulfaguanidine is added to the diets.

---

[9] Chen, K. K., Robert C. Anderson, Frank A. Steldt, & C. A. Mills. Environmental temperature and drug action in mice. J. Pharmacol. and Exper. Therapeutics 79: 2. Oct. 1943.

[10] Mills, C. A. Environmental temperatures and thiamine requirements. Amer. J. Physiology 133: 3. July 1941.

Choline is also required in larger amounts in hot-room diets if optimal growth and development are to be attained. Rats at 70° F. do well on as little as 0.5 gram of choline per kilogram of diet—in fact, they grow normally on a choline-free diet after the danger period for acute hemorrhagic nephritis has been passed. In the heat, however, rate of growth is directly related to the amount of dietary choline, being optimal only at 5 grams per kilogram of food. This substance thus becomes an important growth factor in hot environments.[11]

On a Purina chow diet, growth of rats and mice exhibits a 40% retardation in the heat. This handicap is reduced by half when the hot-room diets are properly enriched with thiamine. Further growth improvement with choline addition leaves only a 5–10% growth deficit in the heat. This remaining handicap can be completely eradicated by raising the protein content of the hot-room diets from 18% to 24% (casein) or by the addition of 0.2% cystine.

After the indicated thiamine, choline, and protein corrections have been made, growth and development in the heat are identical with those seen in cool surroundings. Ear and tail overgrowth no longer takes place in the heat, indicating the likelihood of a greater metabolic efficiency.[12] Body phagocytes from heat-adapted animals are just as active in ingesting bacteria as are those from animals kept in the cold, provided the higher nutritional requirements of the heat have been met.[13] No study has yet been made to see whether such dietary corrections would eliminate the differences in neoplastic tendencies shown by mice of susceptible strains in heat and cold. Nor have experiments been completed to show whether these corrections would make resistance to infection the same at all environmental temperatures.

GENERAL DEDUCTIONS

It now seems obvious that temperature control in animal environments is essential if experimental results are to be uniform and reproducible. This is particularly true where warm-blooded animals are to be used for drug standardization tests, for vitamin assays, or for the determination of nutritional requirements. It is also true in any studies on animal psychology or mental behavior patterns. This factor should be carefully controlled in all studies on resistance to infection,

[11] ———, **C. A.** Environmental temperatures and B-vitamin requirements. Archives of Biochemistry 1: 1. Oct. 1942.

[12] **Mills, C. A.** The B-vitamins in tropical nutrition. Internat. J. Leprosy **10:** 82. 1942.

[13] **Cottingham, Esther, & C. A. Mills.** Influence of environmental temperature and vitamin-deficiency upon phagocytic functions. J. Immun. **47:** 6. Dec. 1943.

red cell and hemoglobin production, and phagocytic functions. And, of course, it is of paramount importance for the maintenance of optimal fertility in breeding stock and of high vitality and rapid growth of the young.

When it is found necessary to grow or use experimental animals in summer or tropical heat without artificial cooling, proper fortification of the diet should be made to meet the heightened requirements. These have been indicated for the rat and are now being determined for the chick. No such studies have been made for guinea pigs, rabbits, dogs or other experimental animals.

Domestic livestock show very clearly the retarding effects of tropical heat, for steers on the best ranches of tropical lowlands take 4–5 years to reach the 1,000-pound slaughter size, whereas this size is attained in 1½ years in our northern states. Hogs in Panama require 15 months to reach the 200-pound size achieved in 6–7 months in cooler lands. Cows of excellent stock, imported from the United States into the Canal Zone and fed the best known food mixtures, produce 25% less milk than they did in cooler climates.[14] The matter is a very important one in animal husbandry and will repay the animal grower for his careful attention to the subject.

Best temperatures for breeding rooms containing suckling young are in the neighborhood of 76–78° F., but the breeders, before delivery, will have highest vitality to transmit to the offspring if kept at temperatures of 68–72° F.; and the young, after weaning, will also have highest vitality at these temperatures.

Uniformity of temperature is of great importance, at all times, in the prevention of pneumonic infections. One series of cancer mice was observed over a 20-month period, under 3 different temperature conditions: 68° F., 90° F., and in the variable temperatures of an ordinary laboratory room. Among 66 mice kept in the air-conditioned room at 90° F., none of those dying showed evidences of pneumonia at autopsy. In the room kept at 68° F., there was one such, but among the 66 exposed to the more variable laboratory conditions, there were 13 with evidence of pneumonia at death.

It will pay animal growers to keep in mind, at all times, that proper ease of body-heat loss is a dominant factor in warm-blooded individuals of any species, that such individuals are combustion machines of relatively low working efficiency, needing quick dissipation of waste heat from the body. Difficulty in such dissipation, if prolonged for

[14] Mills, C. A. Climate Makes the Man. Harper Bros., New York. 1942.

2–3 weeks below the fever-producing level, results in a more sluggish cellular metabolism, a general lowering of vitality, and a drop of fertility.

DISCUSSION OF THE PAPER

In answer to various questions, **Dr. Mills** made the following comments:

The food factor is a very important one, of course. Corn-fed steers mature faster than those ranging on the best pastures Alfalfa is rich in the B-vitamins and greatly accelerates growth when mixed with grain-feeding. On tropical ranches, the grass is of poor quality, reedy and tough, and the steers get no grain. This inadequate diet undoubtedly plays a large part in the retarded development seen there. But we now know that there would be some retardation by the heat, even with the best feeding methods, unless artificial supplementation of the diet is carried out.

Animals in the cold can tolerate the doubled dietary thiamine needed by those in the heat, but they do not do well with the 6-fold increase in choline needed by the hot-room animals.

Oxygen consumption studies should be carried out at a uniform neutral temperature. 28° C. is commonly used and is considered to be the point of thermal neutrality for most warm-blooded animals.

Obviously, oxygen consumption tests carried out at 32° C. could not be compared with those made at 20° C., for the immediately prevailing rate of body-heat loss in the cold would bring on a direct combustion increase (largely through increased skeletal muscle tone). The effect of climate, or of long-continued heat or cold, is of slow onset, requiring about three weeks for its full development. Three weeks of tropical moist heat depresses the metabolism of animals and this depression is still in evidence after a day spent at 28° C. Long-continued cold likewise brings on a metabolic stimulation which persists for several days when the animal is placed at 28° C. It is this slower and more lasting metabolic effect through which climate exercises its chief influence.

In the drug-toxicity studies described in the paper, the animals were adapted for only one week before the injections were made. This work should be repeated with the full 3-week period for adaptation

Even with the most perfect food mixtures so far devised, milk production at the Minde Dairy in the Canal Zone is 25 per cent less than was yielded by the same cattle before being shipped down to the Zone.

ANIMAL COLONY MAINTENANCE—FINANCING AND BUDGETING—VIEWPOINT OF THE UNIVERSITY

By Sidney Farber

Harvard Medical School, Boston, Mass.

The extraordinary advances in the biologic sciences during the past fifty years have been based in great part upon data obtained from the use of laboratory animals. It is recognized, too, that a large proportion of the important progress in human and veterinary medicine would have been impossible without the opportunity for unhampered animal experimentation. The universities and, specifically, the medical and allied schools, therefore, have a real interest in the problems of animal colony maintenance and in a discussion of such considerations of importance to the raising of animals, as genetic purity, nutrition, environmental conditions, and infectious disease. A total of many thousands of mice, rats, guinea pigs, rabbits, cats, dogs, monkeys, hogs, fowl, frogs, turtles and leeches are used each year in the teaching of students and in the conduct of research at Harvard Medical School, the Public Health School, and the School of Dental Medicine. The conditions which govern the care and use of these animals are the responsibility of an animal committee composed of faculty members appointed annually by the President. It may be stated here that all teaching exercises and experiments, involving the use of animals, must be conducted according to regulations set, and under the possibility of unannounced inspection by this committee. Copies of the regulations are displayed prominently in all animal rooms of the school. They guarantee humane and intelligent handling of animals. They are uniform with those adopted by the American Medical Association. For the efficient and economic purchase and delivery of so many animals of such different kinds, the school maintains a central Animal House. This is a self-contained one-story building structurally attached to the school building. It contains a series of rooms designed to hold cages for animals and tanks for turtles and frogs, a feed room, a utility room, a kitchen with facilities for the preparation of foods of various kinds, a large refrigerator and an incinerator. The Animal House is cared for by two full-time men—the manager and his assistant—who order, receive and deliver animals, prepare food, clean cages, care for animals and, in general, keep the quarters in a satisfactory condition. The part-time services of a secretary in the Dean's office are required for the bookkeeping. All

financial transactions are controlled by the Bursar's Office at Harvard University.

The Animal House is, in effect, a retail store for the sale of animals, and of foods such as horse meat, to the various departments of the school. Individual research workers or department heads may deal directly with commercial firms for the purchase of animals without the intervention of the Animal House. Years of experience with this retail store have proved its value. An attempt is made to maintain a stock of animals sufficiently large to care for the current needs of the school with some allowance for unexpected demands. Efficiency is achieved by the cooperation of department heads, who give adequate warning to the Animal House of unusual needs, and who, when possible, place in advance and by date their orders for animals used for teaching or for research projects. The manager of the Animal House has expert knowledge concerning sources of animals, the habits of dealers, care and feeding of animals and other data which he makes available to research workers in the school. The Animal House is open to inspection by any interested person who may visit without warning. The employment of responsible men of good character as permanent workers in the Animal House, the strict enforcement of the rule which permits the purchase of animals only from reputable dealers and the high standards of humane care given to all animals have served to win the approval of the anti-vivisection groups for our methods of handling animals, if not for the purposes to which we put them.

Housing of Animals

Animals may be housed in the animal room or quarters of a given department of the school or in one of two central animal houses which have been built for the purpose as structural additions to the school. Departmental animal rooms vary from simple laboratories converted for this purpose for temporary convenience to more elaborate and permanent units attached to certain departments of the medical school. Similar animal rooms and quarters are found in the teaching hospitals associated with the medical school. Of the two central animal houses, one, a recent acquisition, is designed particularly for the care of animals used in the study of infectious disease. Its basic plan is that of a cooperative apartment house with completely independent units cared for by the departments using the units. The second central animal house is the one which contains the retail store for the sale of animals. This has an additional function, that of a hotel, in the renting of cage space.

Entire rooms or portions of rooms may be leased to research workers who, themselves, provide food and care. If it is desired to have animals cared for and fed in the central animal house, a set charge is made which includes all care, food, overhead and labor. The charge for boarding animals is computed on a cost basis. At the present time, the following charges are made for boarding animals (including room, food, and care):

TABLE 1

Mice	20¢ per dozen per day
Rats	20¢ per dozen per day
Guinea Pigs	2¢ per day
Rabbits	5¢ per day
Monkeys	22¢–28¢ per day
Chickens	5¢ per day
Cats	15¢ per day
Dogs	40¢ per day

Cost of Animals

The charge for animals, made by the Animal House to the research worker, is based on the formula of cost, plus 10 per cent. This surcharge is omitted, in special cases, if no expense to the Animal House is entailed in the transaction. The amount of surcharge varies somewhat from time to time. It is so calculated that the books of the Animal House will show neither loss nor profit at the end of the year. The medical school furnishes heat, light, water, and the building. All other expenditures of the Animal House are defrayed by this surcharge or by the fee for the rental of rooms or cages in the central Animal House. Some variation in the cost of animals is occasioned by the loss of stock, either from deterioration in transit or from infectious disease after arrival in the Animal House. An effort is made to distribute the cost of such loss as evenly as possible, so that no one department will suffer unduly, if there is an epidemic with a high mortality rate among the animals used particularly by that department. Recently, approximately 20 per cent of the cats used in one year died shortly after arrival at the Animal House. The existence of a retail store in the Animal House permitted that loss to be absorbed without upsetting, to a serious degree, the budgets of these department or research workers using cats. The formula of cost, plus 10 per cent surcharge, which is now in effect, brings the total cost to the research worker of the various animals handled by the Animal House to the following levels:

TABLE 2

Mice	18¢ to 23¢
Rats	55¢ to 75¢
Guinea Pigs	85¢ to $1.25
Rabbits	35¢ per pound
Monkeys	$30.00 (subject to change)
Chickens	about $2.50
Cats	$ 1.50 to $3.00
Dogs	$ 3.50 to $6.00

Method of Paying for Animals

There are roughly three sources of funds which are used to pay for animals at Harvard Medical School: (1) the departmental budget, obtained from university funds and assigned by the Administration of the University; (2) special research grants, obtained either from funds entrusted to the University for special purposes, or from foundations interested in the support of research; and (3) the research worker himself. Payment by the individual is not a common practice and is encountered only when a clinician conducts occasional part-time research or performs diagnostic tests on animals for his own benefit. The amount of money obtained from the departmental budget, or from sources outside of the regularly assigned funds to one department, varies according to the department, the character of the projects being investigated, and the size of the project. In general, departmental budgets care for animals used for teaching purposes and for a limited number of experimental animals. Most of the large projects which make use of great numbers of animals over a long period of time are supported by funds not ordinarily provided for by the departmental budget.

University Breeding Farms

The Harvard Medical School conducts no breeding farm as a part of the Animal House. Individual departments may raise mice or rats or guinea pigs, if a research project necessitates the control of factors which cannot be controlled when the animals are procured from commercial dealers through the Animal House. A summarizing statement may be made that it has been found more advantageous to purchase from reputable dealers most of the animals used than to raise them on our premises. The reasons are many. We may mention the amount of space which would be required and the special equipment. Large stocks of mice, rats, guinea pigs, rabbits, etc., far beyond the actual number of animals used, would have to be kept on hand to meet the

demands for animals of a specified age, weight, sex and breed and to offset the depredations caused by epidemics. To maintain such large colonies for the production of the animals needed, great expense would be incurred. This would have to be met in one of two ways: either the medical school breeding farm would have to seek markets outside of the medical school to take care of the surplus animals, and so become, in fact, a commercial breeder, or the cost would have to be absorbed by the medical school and individual departments. This would increase considerably the cost of research. These statements apply to animals used for most of the research work at the Harvard Medical School. Occasional projects, such as experiments concerned with certain aspects of cancer and genetics, demand the maintenance of large colonies of mice, but here the maintenance of the colonies is a part of the research project itself. When large numbers of mice of a specified age and rigidly controlled environmental conditions are required for infectious disease studies, it is necessary to maintain breeding colonies. At the present time, the surplus from one such colony is sold at the rate of 15¢ per mouse, a figure which, under favorable conditions, covers overhead, food, care, depreciation of equipment used, and, perhaps, a small profit. Occasional epidemics in such colonies may distort these calculations considerably and cause a loss which has to be borne by the department itself. The present practice of the Animal House is to raise no animals and to place reliance upon reputable commercial breeders. Of these, there is a sufficient number. Much of the responsibility for the research supply of some of the animals used in medical schools, as, for example, the rat or the mouse, has been lifted from medical schools by the existence of research institutions such as The Wistar Institute and the Jackson Memorial Laboratory. Such institutions, by their contributions to our knowledge of the biology of the rat and the mouse, and by the development of strains of these animals designed for specific purposes, have provided medical research with an invaluable background for the use of these animals in the solution of problems in human medicine.

The Dog and the Cat

All mammals required for teaching or for research, except the cat and the dog, may be procured without too much difficulty, under ordinary conditions, either from commercial breeding farms or from importers, as is the case with monkeys. No commercial breeder now produces dogs and cats for use in research or teaching. The situation

of the medical schools in regard to the supply of dogs and cats is a serious one. In Boston, all cats and dogs are supplied either by a small number of licensed dealers in the vicinity of Boston, or by dealers in Pennsylvania, New York, Delaware and New Jersey. Animals are purchased only from licensed dealers who meet the requirements of local governmental agencies. There is no traffic with boys or with private individuals. It should be noted that all dogs and cats purchased have been discarded by their owners for one reason or another. Research workers, therefore, are compelled to accept, for their purposes, dogs and cats of uncertain age, of unknown genetic constitution and of variable past history, as far as health and treatment are concerned. Frequently, these animals are ill or must recover from the rigors of a long train or auto journey, which is a hardship even under the most humane conditions. The cost of cats varies from $1.50 to $3.00, and of dogs from $3.50 to $6.00 per animal. Included in this cost, is the expense of transportation. In Boston, approximately 1500 dogs and 2000 cats are used each year in the medical schools and research laboratories. In the same city, the animal leagues and humane agencies, acting as animal pounds for the city of Boston, destroy by electrocution approximately 35,000 dogs and 65,000 cats a year. In Chicago, St. Louis, Houston, Durham, No. Carolina, and one or two other cities in the country, dogs may be obtained for the purposes of teaching or research from the city pound under the provisions of a city ordinance. The cost to the medical schools and laboratories is merely that of handling the animals. It varies from $1.10 in Chicago to $1.63 in St. Louis. The cost to the city for the maintenance of the pound or the gathering of stray animals is not included in these figures. Approximately, 10,000 dogs per year are used by the medical schools of Chicago for scientific purposes.

The uncertainty of the supply of dogs and cats in most cities of the country, the great variation in the condition of the animals, and the lack of information concerning their genetic constitution and health background have been responsible for repeated attempts in several parts of the country to raise dogs and cats for experimental purposes. Some of these projects have been suggested by men who have had long experience in the raising of mice and rats, and who are disturbed by the reliance of medical workers on data obtained from the use of dogs and cats of such uncertain background. Estimates concerning the cost of raising a dog vary greatly in different parts of the country. For one attempt, it was learned that it cost about $15.00 to

raise a puppy to the age of six months, a figure which does not include the cost of maintenance and services and the investment for buildings and equipment. Under university conditions in the south, the cost of feeding and caring for one dog for one month was found to be $3.00. This figure includes food, labor, and bedding, but not the overhead of equipment and maintenance. Nor does this figure indicate the cost of breeding animals and raising puppies to adult life, with all of the additional expenses implied by such a project and all of the loss which might be incurred as a consequence of infectious disease and epidemics. Estimates by different workers range from $25.00 to $75.00 as the cost of breeding and raising one dog to maturity. It is obvious that these estimates are not based upon the same data. The larger figure is probably nearer the cost, if buildings and equipment are included.

There are many workers in medicine, today, who favor the use of pedigreed dogs for research purposes, and believe that a great opportunity for research might be provided, as a by-product of such a project, by the availability for study of large numbers of dogs produced under controlled conditions from selected stock. It appears clear that the expense of raising such dogs for teaching and experimental purposes must exceed, to a very great extent, the cost of $3.00 to $6.00, when the animals are procured from licensed dealers, or the even lower cost, when the animals are obtained from the city pound, as in St. Louis and Chicago. No data are available concerning the large-scale breeding of cats for experimental purposes. A constant supply of dogs and cats, raised for the purpose, under conditions comparable to those which govern the raising of mice, rats, rabbits and guinea pigs, and free from the activities of anti-vivisectionists, represents a goal, the attainment of which is surely worthy of great effort. Some solution, not now apparent, must be discovered, however, to bring the cost of commercially bred dogs and cats within the reach of university budgets and foundation grants.

ANIMAL COLONY MAINTENANCE—FINANCING AND BUDGETING; VIEW-POINT OF THE COMMERCIAL BREEDER

By C. N. W. Cumming and F. G. Carnochan

Carworth Farms, Inc., New City, N. Y.

It is an old and time-honored statement that supply is regulated by demand and that, when a demand exists, sooner or later a supply will come into being. The business of raising laboratory animals came into existence to supply the demand for test animals which was caused by the increased activity of medical research, and the necessity of using animals to test the new biological agents coming into use. The rapid growth of experimental medicine since the last world war has encouraged the growth of a firm such as ours, whose entire activities center around the production of uniformly high quality laboratory animals.

In the dark ages of experimental medicine, any type of laboratory animal was difficult to obtain and many an experimenter was limited in his work because of this shortage. Two palliatives were resorted to in order to overcome this condition. Some of those who could command the necessary facilities, and who had sufficiently large budgets, started to raise their own animals. In this discussion, we will call them consumer breeders. There were others, who either did not wish to enlarge the scope of their activities to include animal husbandry, or who did not have the necessary funds available. These took the easier way of encouraging individuals in the vicinity, whom we shall call home breeders, to raise the particular type of animal that was required.

Consumer Breeder

Among the consumer breeders, were many pharmaceutical houses who kept accurate records of their costs. In most instances, after extensive experience, these costs proved excessive and the operations were discontinued. These high costs were, in large part, due to the fact that it was impossible to correlate production with demand. This resulted in an excess of animals, which had to be destroyed when they were not required, and in a shortage in times of great need. This type of consumer breeder, therefore, turned to the commercial breeder as a dependable source of animals bred to specific standards at a reasonable price. Consumer breeders that continue to raise animals, regardless

of cost, are those institutions which produce only a small part of their requirements; those few which are endowed with unlimited budgets; and some commercial houses that are compelled to do so by government regulations.

Home Breeder

Those who had chosen the encouragement of home breeders, as a means of solving their problem, also found themselves in difficulties in so far as supply was concerned. This was due to unforeseen increases in their requirements, as well as to the difficulty of estimating future demands. The small breeder, on the other hand, who started to raise animals for some particular laboratory, soon found himself in trouble, partly, from his own inability to look on his newborn enterprise as a business, and partly, from the very uncertain demand for his animals The home breeder is usually engaged in breeding as a means of added income, using waste space which is frequently inadequate and unsuitable for the purpose. An additional disadvantage is his lack of basic biological training necessary for the scientific breeding of specific types.

Many of these new breeders have started to raise animals with the high hope that they would not have any trouble marketing their product, but have found themselves left high and dry, in spite of an overall increase in demand. This condition has obtained because the small breeder has been dependent for his sales on one local laboratory or institution, in which the demand has, in accordance with the very nature or research work, varied considerably. If a research project, which necessitated the regular use of a large number of animals, is suddenly concluded, and no other work is started immediately, the local breeder may be without his market for a matter of months. After a few futile attempts to sell his animals elsewhere, he usually decides to go out of business. On the other hand, if he is lucky enough to find another laboratory to use his animals, the agony is only prolonged, since, sooner or later, that project ends and he again has no market for his animals. He then is faced with the necessity of either killing his stock or going bankrupt feeding them. Furthermore, he has no means of determining actual costs, since, by ignoring the hidden costs, he generally considers his profit as the difference between his sales receipts and his direct expenses. Eventually, he finds the return inadequate, and, with better opportunities in other directions, he gives up the venture. For these reasons, consumers have generally found that the

home breeder is not a dependable source of standard animals and have turned to the large commercial breeder for their requirements.

Commercial Breeder

Because the commercial breeder must operate on an extensive scale, he must provide adequate capital for his needs. Also, the diversified nature of the management of such a business requires that several men, with capabilities to cope with the varied problems, join in the enterprise. It is clear, therefore, that the corporate structure is ideally suited to meet these conditions. This requires the organization of a corporation, with its attendant expense, as well as the sale of stock to qualified men for sufficient capital to do business.

While this capital investment meets the basic needs, an examination of the requirements of a well-run integrated business will show the need of additional outside capital. This may be secured by means of mortgages on real estate and chattels, bank loans and by indebtedness to creditors for purchases made.

These varied forms of invested capital will be evidenced by the following assets on the books of the company.

Cash. A sufficient balance of cash must be maintained in the bank account to meet the current needs of the business, such as weekly payroll, purchases of materials and supplies, and operating expenses, such as electric light and power, rent and telephone.

Accounts Receivable. Since sales are generally made on the basis of credit, and payments, therefore, are not received until the lapse of thirty days or longer, sufficient capital must be provided to carry such accounts until payment is received. This amount will be considerable as it may equal one sixth of the annual sales.

Inventory. In order to have stock for sale, a breeder herd must be maintained. Since this herd is comprised of specially selected animals which must be kept for a considerable time in reserve before they become satisfactory replacements for dead or obsolete animals, the cost of such animals is determined to be equal to five times the cost of animals held for sale. In the case of mice, this breeding herd must be reconstitued on a seven- or eight-month basis depending on the strain. That is, no mouse is permitted to remain in the breeding colony for a longer period. Our experience has shown that the quality and quantity of the progeny beyond that point does not warrant the cost of maintenance.

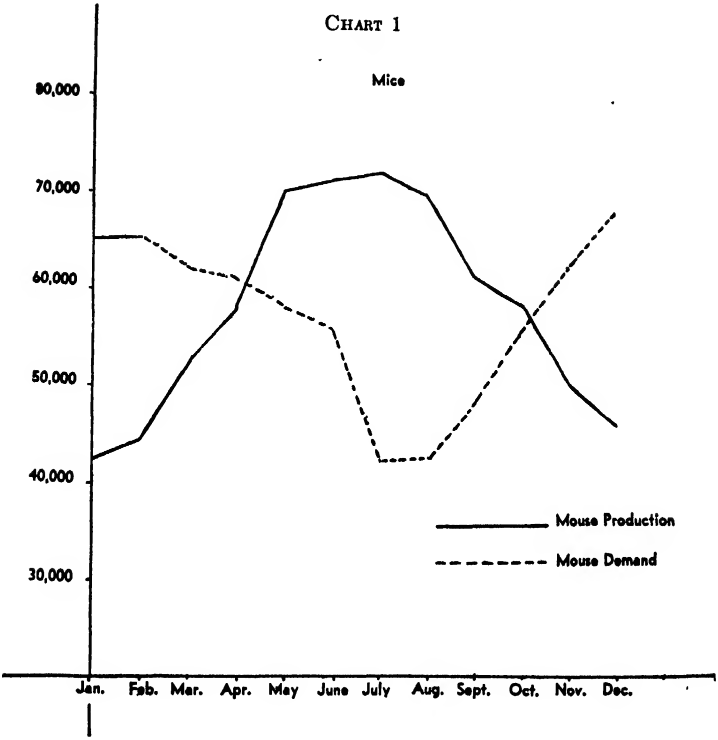
The size of the breeding herd is dependent upon the forecast of future need for animals. It is estimated that, to increase the sales stock of mice by fifty per cent for a period nine months hence, it is necessary that the breeding herd be increased ten per cent per month for five months. To increase the breeders by ten per cent, it is necessary to introduce a gross total of twelve per cent. The period of nine months, mentioned above, is the average time required to complete a program of expansion of any size, provided the sales stock is kept at a level sufficient to ensure income for necessary operations. In this way, a sufficient sales stock of animals of different sizes and strains is on hand, at all times, to supply, at least, the most urgent demands during expansion. Incidentally, the job of attempting a selection of the most urgent orders is a thankless one, because invariably every order is considered a rush order by the consumer, who deems it a personal affront if his requirements are not met at once. The demand for animals cannot be satisfied on such short notice as orders for manufactured goods, but must be met by production estimates long in advance of anticipated sales. An important factor to keep in mind in this connection is that, as a rule, the period of greatest demand does not coincide with the period of greatest production (CHART 1). It is, therefore, necessary to balance all factors in planning the production, since an oversupply results in losses in two directions. The cost of feeding and servicing the animals is lost if they are not sold, and the animals must be destroyed when their age makes them unsalable. Then, too, the overcrowding of the cages make the possibility of disease an ever-present danger.

With these factors in mind, our experience has shown that the inventory of mice breeders on hand should, at an average, equal approximately two-thirds times the anticipated weekly sales; and, for guinea pig breeders, six times the anticipated weekly sales.

CAPITAL EQUIPMENT. The business of raising animals as a commercial venture necessitates investment in many kinds of capital equipment. Most of these items are of a type peculiar to this business. The buildings should be erected specially for this purpose and should contain many essential features, such as air-conditioning, vermin control and disease barriers. They should be designed for efficient servicing, feeding and cleaning. The grounds must be spacious and provide sufficient arable soil to raise part of the required animal feed.

The equipment of the building is of two types—fixed and operating. The fixed types include heating, plumbing, sanitation, air conditioning and animal [illegible]es. The operating equipment includes such articles

CHART 1



as animal cages, cleaning equipment, bottles and feeding wagons. The same furniture and office equipment, necessary for any other business enterprise, is required by the commercial breeder, which includes desks, typewriters, safes, calculating and adding machines.

AUTOMOTIVE EQUIPMENT. Trucks must be provided for transportation between the different colony units and for other internal hauling work. Trucks are also required for deliveries to local customers, to freight and express stations, and for the pickup of incoming supplies.

Besides the above mentioned items of capital investment, it is found that some capital is tied up in other miscellaneous assets such as security deposits, prepaid insurance, loans and advances to employees.

Price Determination

One of the most important functions of management is the determination of the price of the product, because, if it is too high, the product is

unsalable, and, if too low, the result is a loss. Another important function is the constant effort to reduce costs so that the selling price can also be reduced.

In determining the selling price, all the factors of cost must be considered (CHART 2). The direct costs include feed and labor, operating supplies and delivery expense. The indirect costs include supervision, laboratory expense, depreciation of equipment and repairs. The administrative costs must also be added and include executive salaries, office salaries, telephone and telegraph, postage, printing and stationery, business and property taxes, insurance, and interest on borrowed capital.

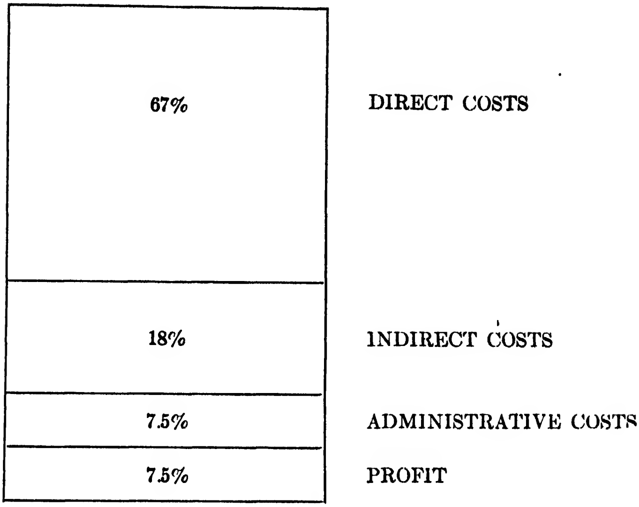
In arriving at the selling price, the total of these costs must be divided by the salable production, which gives the unit cost, to which must be added a profit. The factors which affect costs in this business are different from those met with in most businesses. A machine shop owner, spending a dollar in production, knows what his output will be within a very narrow margin. In this case, when a dollar is spent for production, the results can be estimated only on a widely fluctuating basis. The production for any period is not necessarily in direct proportion to the amount spent, but is affected by losses due to climatic fluctuations, disease, and other uncontrollable factors.

Many of the foregoing items of direct cost are apparent and are probably taken into consideration by all breeders. It is when we pass on the indirect costs that the major differences between apparent and real costs are encountered. It is the failure on the part of many breeders to take these indirect costs into account that encourages them to offer stock at lower prices, which results in the very high percentage of business failures in this industry.

Budgeting

Good management works on the basis of a budget prepared for the future on the experience of the past and providing for the probable demands. The budget must be based on the sales which are planned as probable, and the costs required to fill that demand. The costs must include all of the elements set forth below as aggregating the cost of operation. Provision for financing the budget must be planned so that the funds will be available when needed. If all the elements of the budget work out as anticipated, and efficient operation keeps the costs within the bounds set, then a profit on operations results. Failure

CHART 2



in any particular, or the occurrence of an unexpected calamity not covered by insurance, results in a loss.

The main elements of cost as planned in our budget are as follows:

A. DIRECT COST (67 per cent)

1. *Feed:*

This item represents about one-third of our direct costs. Our practical research, comprising actual feed testing programs, has assured us of the best possible results for this important factor in animal husbandry.

2. *Direct Labor:*

This comprises about one-half of our direct costs, consisting of wages paid to animal tenders. All labor employed for this purpose is specially trained to perform every operation in the most efficient manner. These methods have been perfected as a result of detailed time studies.

3. *Delivery Costs:*

4. *Operating Supplies:*

Items three and four account for the remaining one-sixth of our direct costs. Delivery charges are determined by Railway Express Company rates, over which we have no control. Another item is the cost of shipping containers. These containers are used only once, due

to the possibility of contamination from wild rodents en route. The budget for Operating Supplies is based on past experience and is kept as low as is compatible with efficient operation.

B. Indirect Costs (18 per cent).

These include such accounts as supervision, laboratory expense, rent, electric light and power, operating expense, depreciation of buildings and equipment, repairs and indirect labor. Most of the above accounts are common to all businesses. However, the item of laboratory expense is not an operation that is commonly incurred in the laboratory animal breeding field. It includes the cost of testing the breeding herds for most of the common animal pathogens. Supervision is also noteworthy, for its covers a multitude of very necessary functions and controls besides that implied in the name. The packing and final inspection of all our animals is done by the supervisory force and, on the excellence of this work, to a large extent, depends our reputation. Also, under this heading, comes the keeping of all production records and the performance of genetic tests to maintain the purity and uniformity of our strains.

C. Administrative Cost (7.5 per cent).

This comprises the following expenses: executive salaries, office salaries, telephone and telegraph, postage, printing and stationery, legal and accounting, business and property taxes, social security taxes, insurance and interest on borrowed capital as well as miscellaneous expenses.

D. Profits (7.5 per cent).

Although our budget is planned for a profit of seven and one-half per cent of the selling price, our past experience has shown that unforeseen contingencies arise which may absorb a large part of this item Furthermore, after making provision for expansion, it has been found that the capital invested has shown no return to the stockholders in the form of profits. The principals of the business draw a modest salary and are content in knowing that their contribution is a factor in the progress of medical research.

Discussion

The commercial breeder is able to command a wide and varied market for his animals. This is possible by reason of the fact that the company employs the services of at least two executive officers, one of whom looks after production and research, while the other confines himself to sales and supervision of the office. Both of these officers

must coordinate their efforts to formulate the general policies which guide the company. At the same time, each must devote his energies to the departments under his guidance. In this respect, the commercial breeder has a definite advantage over the organization consisting of only one man, whose efforts are necessarily so divided as to cause detriment to either sales or production.

It is this ability to command a large market that enables a commercial breeder to produce large quantities of animals. This, in turn, makes it possible to maintain an inventory sufficient to supply, not only large regular orders, but also most emergency orders. The fact that the commercial breeder has proven his ability to stay in business, year after year, makes possible a constant source of supply. Because this supply is maintained, in so far as is possible, under standard conditions of environment and heredity, it enables the research worker to duplicate, not only his own experiments, but also those of other investigators who have used similar uniform strains.

The large scale of the commercial breeders' operations makes it possible to institute research programs, whereby strains are produced for specific purposes. The use of these specific strains makes it possible for the investigator to obtain satisfactory results with fewer animals, thus reducing the expense of his experiment, even though the unit cost of the animals may be slightly higher. Our company is eagerly awaiting the day when it will be possible to resume research dealing with animal diseases to which our own colonies may be susceptible. This research has been discontinued, due to the pressure of war orders, but plans for new investigations are well advanced, and it is anticipated that the near future will find them in progress.

The cost of animals is kept as low as is consistent with good business practice, and our company feels that a very necessary part of this cost is the maintenance of an efficient office staff. This enables correspondence to be dealt with promptly, and accurate records to be kept of all customer transactions, in addition to the usual business operations. We have found that the keeping of accurate sales information has enabled us to build up a fund of knowledge about our various strains which makes it possible for us to advise the research worker as to the type of animal most suitable for his particular problem.

The large commercial breeder of laboratory animals has been of material assistance to both government research programs and manufacturers of biologics during this war. His large production has gone a long way towards stabilizing his prices without any policing from the

O.P.A. During the last war, there was no organized business in this field, and, since the home breeder abandoned his colony to enter industry at a high wage, the shortage of mice was sufficient to drive the price to one dollar per mouse. During this conflict, the same condition applies, as far as the home breeder is concerned, but the large commercial breeder, since he expects to continue in business after the war, advances his prices only to the amount necessary to cover extra war costs. To our knowledge, the highest price charged for mice by a reputable breeder, during this emergency, has been 25¢ each.

Without overstatement, it may be said that the commercial breeder performs a vital function by supplying a suitable medium for testing medical theories and biological products.

ADDENDUM

Because of the great interest shown in the paper on heating, we thought this group might be interested in our experiences with heat controls.

In the first building which we erected, we installed a direct hot-air system with a blower behind it to force the air as evenly as possible to all parts of the building. This was controlled very simply by an electric thermostat which turned the oil burner off and on as heat was required. We have found many disadvantages in this system The heat would fall very rapidly in certain sections before the thermostat could catch up with the temperature changes. The thermostat itself, because of the large amount of dust in the air of an animal house from the bedding and feed, became very irregular in its operations, and the fluctuation in temperature was very wide because of the stickiness of the contact in the thermostat.

In our second building, we adopted a much more elaborate system in which the air is blown over steam coils which have automatic valves pneumatically controlled. The thermostats which are set in the animal rooms are pneumatic also. The boiler is kept under steam pressure at all times, hence, responses to the thermostat are very quickly made and the temperature fluctuations are held within a one-degree range. The heated air, which is passed over the steam coils, is positively driven to each room by a powerful fan and there is equally positive exhaust from the room by a separate exhaust fan. As a safety measure, each room contains a thermostat, which pneumatically controls the current to the fans, to shut them off should the temperature exceed 80° in the room. There is always the possibility of a sticky steam valve, or a leak in the

pneumatic system, which would permit the steam valves to open wide, and so overheat the room. Since this 80° thermostat shuts off the fan completely, no excess heat can be driven into the rooms in the event of such failure. In the air-return plenum chamber is a thermostat set for 60° which will also shut off the fans, in the event of a steam failure, and prevent cold outside air from being introduced into the rooms. This system has worked very well, but, since we believed it could be further improved, in our latest building, we have modified it by removing all the thermostats from the rooms and placing them in the return air current. The advantage gained by this is that attendants cannot shift the thermostat up or down. To date, the control of temperature in the rooms under the latter system has been even more accurate then under the former. The installation however, is so new that it has not yet had a chance to be tested under extreme weather conditions.

DISCUSSION OF THE FINANCING AND BUDGETING PAPERS

Dr. Ellis J. Robinson (*American Cyanamid Co., Stamford, Conn.*):

We have been breeding beagles and my figures agree with those given by Doctor Farber, of about 40¢ per day. My figures indicate that it costs about $150 to raise a pup to the age of a year. We have been using these dogs for a certain type of work and have found them to be a very worth-while experimental animal. I have heard opinions to the contrary about dogs.

Dr. Edmond J. Farris (*The Wistar Institute of Anatomy and Biology, Philadelphia, Pa.*):

I have heard that it is a question of the size of the animal and type of maintenance, that determines the variable costs in dog raising, with range from $15 to $45 per year.

Mr. Carnochan:

In 1938, we were asked to produce dogs in quantity for an institution that was working on immunization and distemper. The dogs would have cost $18 and, on that basis, they refused to buy, because they were too expensive

Dr. Myron Gordon (*American Museum of Natural History, New York, N. Y.*):

There should be a clearing house for information from which we can obtain information on certain types of animals. The National Research Council should be able to establish such a clearing house where we can write and get the information we want. I should like to hear some comments on the subject and get some suggestions and have it brought to the attention of the National Research Council.

Dr. Clarence A. Mills (*University of Cincinnati, Cincinnati, O.*):

A problem in getting dogs for experimental purposes is that often it is difficult to obtain information concerning the genetic set-up of the dog. It takes time to get the dog in condition for the experiment. There should be a standard nutritional and metabolic basis for the raising of the animals for this purpose

Dr. Robinson:

It is necessary to bring up the dogs and keep them for several months. Our experiments, done so far, have been of a hematological and histological type. The dogs have been most satisfactory. We have been using beagles and have used dogs of this kind that we know are completely healthy.

Dr. Mills:

Do you worm the dogs?

Dr. Robinson:

We have found a few worms on autopsy. We have maintained a very healthy condition for the dogs and examine them regularly. The problem should be brought to the National Research Council as to where one can obtain such dogs, with the assurance that they will be suitable for our purposes.

Dr. Herbert Clark (*Gorgas Memorial Laboratory, Panama*):

It is difficult to keep the dogs, parasite free. We have been able to produce good dogs even though they have been infected. The food is a very important item, as described here. We have been forced to use such things as rice. We let the dogs range during the day and lock them up at night. Our estimated cost per year is $40. We get native rice. The dogs produce well, and we do not worm them. Unless, of course, they have a severe case of worms.

Mr. Carnochan:

I should like to return to Doctor Gordon's remarks as to information, and where it can be gotten, about animals. I would like to see that section of the National Research Council, which is concerned with animal strains, establish a definition for "strain" so that we all talk about the same thing. At the present time, the word means anything from a haphazard colony which has been held a long time without outside admixture, to a group which has twenty-one or more generations of brother-sister mating behind it. Some trace back two or three generations before a common pair of ancestors can be found; some, ten, twelve, or even more. A scheme of inbreeding should be established, e.g., twenty generations of brother-sister, and from then on, first cousin, or double first cousin matings. These remarks apply to mice, but are equally true for guinea pigs, rabbits and rats. I suggest that no animal deserves the designation "strain," unless theoretically at least 80 per cent homozygous for all characters.

Dr. Gordon:

You must have a clarity of terms to get what you want. There must be some organization where these items can be straightened out so that we can get the animals we want.

Dr. Carl G. Hartman (*University of Illinois, Urbana, Ill.*):

The National Research Council is the logical place to establish such an information center.

Dr. Farris:

During the course of the second paper, question was raised how to increase the supply of animals during the winter season, when the demand was greatest and production lowest. My suggestion would be to modify the lighting conditions in animal colonies during the off-breeding season, to increase production. By simply increasing the length of day up to twelve hours by artificial lighting, there will be a forty per cent production increase.

Dr. Robinson:

About lighting, what do you consider a good proportion for good lighting with only artificial light?

Dr. Farris:

For rodents, 12 hours light and 12 hours dark to standardize the animals.

Dr. Gordon:

We have used light for increasing production in fish, and have increased production during the spawning season by photoproductivity.

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES

VOL. XLVI, ART 2. PAGES 127–152.

JUNE 29, 1945

A HITHERTO UNDEMONSTRATED ZOOGLEAL FORM OF *MYCOBACTERIUM TUBERCULOSIS**

BY

ELEANOR ALEXANDER-JACKSON†

CONTENTS

* Awarded an A. Cressy Morrison Prize in Natural Science in 1944 by The New York Academy of Sciences. Publication made possible through a grant from the income of the Centennial Endowment Fund. The research was made possible by a grant from the Rosenwald Family Association.

† Department of Public Health and Preventive Medicine, Cornell University Medical College, New York City

INTRODUCTION

The existence of non-acid-fast forms of *Mycobacterium tuberculosis* has been recognized at least as far back as 1900, but the reproductive role played by them has been variously interpreted by many observers. At the present time, there are four main schools of thought.

The first school regards non-acid-fast rods and granules as degenerative in nature, since they are to be found in old cultures, and the supporters of this view, represented by Wycoff,[22] Rettger,[21] and Oerskov,[30] claim that they were unable to observe any mode of reproduction, other than binary fission.

The second school includes those who believe in the filterability of certain elements of this organism (Ferran,[38] Fontes,[38] Karwacki,[14] Valtis,[25] Vaudremer,[26] and others).

The third school includes those who, while unable to demonstrate filterable forms, are of the opinion that the non-acid-fast rods and granules are, in many instances, young organisms which have not as yet developed acid-fastness, but later on develop into acid-fast rods. The adherents of this latter group (Kahn,[12] Nonidez,[13] Lurie,[16] Alexander[1]) agree that binary fission may be the common method of reproduction, but find that non-acid-fast granules, obtained through segmentation, either elongate or sprout into the rod forms of *Mycobacterium tuberculosis*

In a critical summary of the many opinions held on non-acid-fast forms of *Mycobacterium tuberculosis,* A. R. Rich,[42] in his recent book, "The Pathogenesis of Tuberculosis," accepts the results and conclusions of Kahn based on his single-cell observations. On the other hand, he states that "As yet there is no proof that, when introduced into the animal body, the granules are able to produce disease, whether by their own action or by giving rise to mature bacilli in the tissues."

C-E. A. Winslow,[43] in an address on "The Changing Bacteria," expresses the opinion that "such work as that of Morton Kahn, in which the entire process of cell-disintegration has been observed under the microscope in single-cell cultures, should convince the most skeptical."

Finally, there is also a group, (Yegian and Porter[37]) that regards non-acid-fast forms, in many instances, as artifacts brought into being by injury suffered by the microtome knife or by manipulative damage with the platinum loop in preparing slides for staining. The conclusions drawn from their work are vitiated by the fact that the authors describe a technic so different from those usually employed in producing stained preparations that no comparison can be made.

PRELIMINARY OBSERVATIONS

In 1934, while studying stained preparations of rooster, guinea-pig and mouse tissues infected with R and S dissociants of avian and human tubercle bacilli, it was noted that, not infrequently, even in smears made from severe lesions, there were scarcely any *Mycobacterium tuberculosis* to be found. A number of non-acid-fast and acid-fast granules were seen, which, while suggestive of "Much"[18] granules, could not be differentiated with any certainty from non-tuberculous cocci. It was felt, however, that such pronounced lesions must surely be accompanied by this organism in some form, possibly in a form other than that of the classical acid-fast rod. This is a viewpoint long held by other workers.

In 1929 and 1936, Kahn,[12] and Kahn and Nonidez[13] published papers in which they offered the first actual proof of the ability of non-acid-fast rods and granules to develop into the characteristic acid-fast rods from which they were derived. Kahn accomplished this by means of a carefully carried out single-cell isolation of the rods in microdroplets so minute that their entire circumference was visible under the oil immersion lens. The organism he employed was a young, rapidly growing culture grown on Long's synthetic medium. In 1936, Kahn and Nonidez[13] sectioned young and old colonies, showing a preponderance of non-acid-fast types in the former.

In 1934, the writer was able to observe a similar course of development of tubercle bacilli which had been lightly seeded on slants of Bordet-Gengou medium. Stained preparations made during the first week of growth invariably showed a marked preponderance of non-acid-fast cocci and rods. After that, the rods became either a mulberry color on staining, or else developed acid-fast bodies within the young rods. On one occasion, a non-acid-fast rod was seen with one red-staining round body and several others of the same nature still staining blue when subjected to the Ziehl-Neelsen staining method.

It was felt, however, that the usual Ziehl-Neelsen counterstains; namely, either methylene blue, Loeffler's methylene blue, or brilliant green, were inadequate for revealing the non-acid-fast forms of the tubercle bacillus. Too often, they appeared pale and indistinct. An attempt was made, therefore, to intensify their resolution so that they would stand out as definitely as the acid-fast forms encountered. This was accomplished by adding six to eight drops of sodium hydroxide of an optimum concentration to slides when they were flooded with Loef-

fler's methylene blue counterstain. This simple expedient, while excellent for smears of pure cultures, was not suited to preparations from actual tuberculosis material, because the background of the slide was stained a too deep blue. Moreover, it did not differentiate between the non-acid-fast rods and granules of *Mycobacterium tuberculosis* and other non-acid-fast organisms which might also be present.

It seemed possible that, since the acid-fast forms had the peculiar ability to retain the red carbol-fuchsin stain in the presence of mineral acids, the non-acid-fast forms might demonstrate some such resistance to decolorization of the strongly alkaline methylene blue in the presence of a reducing agent. Sodium hydrosulfite was found selectively to bleach all of the blue, or most of it, from the background, without affecting the deep blue non-acid-fast forms of *Mycobacterium tuberculosis* The bleach was mild enough so that the structure of tissue cells and other bacteria were not destroyed. To stain these, a third dye, acid green, was briefly added. This triple stain technic is described in an earlier publication.[31]

When the triple staining method was applied to material of tuberculous origin, such as chest fluids, gastric washings, sputum, and pus, it was noticed that another even more deeply blue-staining form was often present in addition to the usual blue-staining rods and granules. When this form first met the eye, it was thought possibly to be an artifact, or else some other species of organism previously unobserved. It consisted of an amorphous substance in which granular elements or larger globoid bodies were embedded or enmeshed. Although these merocyte-like or zoogleal clumps varied in size, there was a marked constancy in their form and structure. Further study of this form revealed that, not infrequently, it was semi-acid-fast or even frankly red in color when stained with Ziehl-Neelsen's method. Such forms, when acid-fast, at least, have doubtless been seen, but considered as lifeless debris. Referring to the literature it was found that, as far back as 1883, Mallassez and Vignal[17] described zoogleal masses as well as isolated micrococci found in tuberculous lesions, both of which they claimed could give rise to rods and characteristic tuberculosis, as is evident from their Conclusion No. 4, quoted verbatim:

"Ces mêmes tuberculoses non bacillaires avec ou sans zoogloées distinctes peuvent donner lieu, après un plus ou moins grand nombre de générations d'inoculation, à des tuberculoses bacillaires; comme si les zoogloées, les microcoques diffus et les bacilles étaient simplement des formes différentes, ou des états de développement différents d'un même micro-organisme. Cependant, cette transformation des zoogloées en bacilles n'ayant pu être encore constatée directement, on n'est pas

en droit d'affirmer que les tuberculoses bacillaire et zoogloéique sont de même nature, quoique ce soit peut-être l'hypothèse la plus vraisemblable."

The drawings of the zoogleal masses were not made under sufficiently high magnification to say whether or not they consisted of individual forms similar to those described here. It is quite possible they did.

The granule elements in the clumps appeared similar in size to "Much" granules, and often free granules and rods grouped in characteristic manner were noted near the zoogleal clumps. It had also occurred to the writer that the clumps might be some sort of streptothrix or fungoid form, but cultures were made on Saboureaud's medium without growth resulting. A distinguished mycologist who examined one of the stained preparations containing zoogleal forms failed to classify them.

HANGING-DROP STUDY OF CHEST FLUID

Preliminary hanging-drop observations of *unstained* material, both chest fluids and pure culture suspensions of tubercle bacilli, indicated that the zoogleal clumps seen in preparations stained by the triple staining method were not artifacts.

An attempt was next made to determine whether they were living organisms or some sort of cell disintegration product. A chest fluid was obtained immediately after withdrawal with aseptic precautions from a patient at Harlem Hospital, diagnosed as an acute early case of tuberculous pleurisy. The fluid, well protected with sterile plug and outer cover, was taken at once to the laboratory, and tested for contaminants by plating some of it out on blood agar. At the same time, a hanging-drop was very carefully made, and sealed on to a hollow ground glass slide with sterile vaseline. Some of the chest fluid was inoculated into the inguinal gland region of a guinea-pig, and a culture was made on Bordet-Gengou medium.

The hanging-drop culture was carefully examined just after being made. No rods were seen, but a group of zoogleal clumps was noted near the edge of the drop. A drawing was made of these, and the preparation was left under the high power objective in the same position for more than two weeks. The entire set-up was incubated at 37° C. An examination was made every few days, and drawings were made.

After three days, the amorphous clumps containing granules became exclusively groups of coccoid bodies, some greenish and refractile, and varying in size. Some of these were free in the medium. After five days, some of the coccoid forms had developed fine hair-like projections reminiscent of those described by Kahn.

After one week, these projections had increased in thickness, and more of them were noted. Two days later, they were even more rod-like. At the end of two weeks, there were a number of young rods. After two and a half weeks, it was decided to stain the drop by the ordinary Ziehl-Neelsen method to see whether or not acid-fast rods had developed. This was accomplished, and examination revealed a few acid-fast, and a number of mulberry colored semi-acid-fast rods. The results of this experiment, while not conclusive, were considered highly suggestive.

After six weeks, the guinea-pig which had been inoculated with some of the chest fluid, developed slightly enlarged inguinal glands, but showed no signs of general infection. Smears made from the glands showed no acid-fast rods. However, a positive culture was obtained from that portion of the chest fluid which had been seeded on Bordet-Gengou medium, and a guinea-pig inoculated with this growth developed a typical generalized tuberculosis. Smears both from the Bordet-Gengou culture and from the second guinea-pig revealed characteristic acid-fast tubercle bacilli.

EXAMINATION OF NON-TUBERCULOUS CONTROLS

Non-tuberculous control material was next stained by the triple stain technic. Normal human and guinea-pig serum, tap water and hay infusion, when stained by the triple stain method and examined, showed deep blue-staining zoogleal forms quite similar to those seen in tuberculous material. Hanging-drop cultures of normal guinea-pig serum containing zoogleal forms were observed. In a few days, the zoogleal forms had developed into diphtheroid rods. A chromogenic non-acid-fast strain of diphtheroids was isolated from the heart's blood of a guinea-pig which had been inoculated with tuberculous chest fluid, but which failed to develop any sign of disease. These diphtheroid rods stained a green color with the triple stain, that is, they were not resistant either to decolorization by acid alcohol or by the sodium hydrosulfite bleach. They were entirely non-pathogenic when inoculated in large amounts intraperitoneally into another guinea-pig. When these rods were placed deep into glycerol lung broth (under microaerophilic conditions), they were observed to lose their rod forms, and "degenerate" into amorphous zoogleal masses.

Finding blue-stained zoogleal forms in non-tuberculous material, caused the writer to consider abandoning the entire problem. However, upon further reflection, it was considered possible that since

Cornybacteria and Mycobacteria are closely related, both diphtheroids and tubercle bacilli might be able to enter a similar stage.

OBSERVATIONS AND EXPERIMENTS WITH PURE CULTURES

Instead of leaving the problem, it was decided to turn to a study of known pure cultures of human tubercle bacilli. Stained preparations of cultures from Bordet-Gengou slants and fluid media (Long's medium and glycerol lung broth), had already revealed zoogleal forms. In fluid cultures, they were noted in number after about two to three weeks in fluid taken from beneath the pellicle of visible growth, and seemed especially numerous in those bottles where the fluid had become opalescent (not turbid). This fluid was always tested for sterility on blood agar or hormone agar at the time of making stained preparations, and sterility tests were placed in the anaerobic jar to test for the possible presence of anaerobes. No growth was obtained. Fluid which contained seemingly no acid-fast rods or granules but myriads of deep blue-staining zoogleal forms, was also obtained from the serous fluid at the bottom of a Bordet-Gengou slant, five days after inoculation with a suspension of human tubercle bacilli strain H3 S. This fluid was inoculated into a guinea-pig which developed a generalized tuberculosis with lesions more of the limited R type than those of the more miliary S type of infection. Smears made from the animal showed acid-fast bacilli, and sections of the liver, lung and spleen showed evidence of tuberculosis.

A careful and intensive hanging-drop study of known *Mycobacterium tuberculosis* was planned and carried out. Several strains were observed, including strain H37, nationally used in research on tuberculosis. The media used were Long's synthetic medium, and a control medium, namely, specially prepared sterile triple distilled water (Eli Lilly), free from any chemical or bacteriological impurity, and employed by the medical profession in intravenous injection. The glassware was cleansed and sterilized, according to recommendations of Schwabacher,[29] in an extensive article on the dangers of contamination by acid-fast saprophytes from air or tap water of material studied for the presence of very small numbers of tubercle bacilli. In carrying out this experiment, the use of tap water was avoided in rinsing test tubes, bottles and slides. The operations were carried on in a "sterile" room well washed down with 5% phenol.

A light suspension of tubercle bacilli was made by adding a small loop of organisms from a four week's old slant culture to a sterile glass bottle containing glass beads and 2 ml. special sterile distilled water, and then carefully shaken by hand. The bottles were corked with thoroughly boiled rubber stoppers. When the suspensions appeared homogeneous, several more mls. of water were added. A few minutes were allowed for larger clumps to settle, or, in other instances, the suspensions were filtered with sterile precautions through two layers of sterile filter paper after the technic of Petroff, used for colony morphology studies.

Clean cover-slips taken from a petri-dish of 95% ethyl alcohol were carefully flamed, set down, and at once covered with a sterile petri-dish cover. Then, with a platinum loop, tiny blobs of sterile vaseline were placed at the outer corners of each coverslip. In the center of the cover-slips were placed, respectively, one small loopful of medium, and then one small loopful of bacillary suspension was mixed with it The coverslips were covered quickly by inverted hollow ground glass slides The small amount of sterile vaseline on the corners of the coverslips was just enough to make them stick tightly when the preparation was turned right side up. A more thorough sealing with sterile vaseline was then made around the rim of the coverslips. Complete preparations were made one at a time so that each drop was enclosed by its sealed space as soon as possible Such hanging-drop preparations could be kept for months. Control drops of the media alone were also made. Each preparation was placed on moist filter paper in a glass petri-dish, and incubated at 37° C., after a first examination had been made under the high power objective of the microscope.

Careful daily observations were made, and the following changes were revealed:

1. The rods tended to swell.
2. Granules appeared in a number of them.
3. After a week, some groups of rods had broken down into a plasmodium or amorphous material containing granules.
4. After two weeks, few rods were seen; some granules were free in the medium, some were surrounded by amorphous substance
5. No further change was observed under these conditions

A moving picture record was taken of the changes undergone by a group of *Mycobacterium tuberculosis* rods in a hanging-drop culture in Long's medium. A single clump of rods at the edge of the drop was observed continuously by training the high power objective of the microscope on this clump, and taking pictures at intervals of several days, or whenever any change was noted. Detailed drawings were also

made. Continuity was established by observing changes from rods to zoogleal forms in a single clump of organisms.

When no further change was observed, an attempt was made to induce, if possible, the reversion of the zoogleal forms to rod forms, by adding a tiny loopful of glycerol lung broth to the same drop with aseptic precautions. Although continuity was necessarily lost with respect to the single clump which had been under observation, changes were noted developing within the drop, and these were recorded also by the motion picture camera.

The first change in the drop after lung broth had been added, was an increase in the number of *free* granule forms. These were of varying sizes. Again, as had been seen earlier in the chest fluid hanging-drop, after five days, tail-like extensions appeared on some of the granules. These delicate tails thickened into rods. In some cases, the rods seemed to form by a lengthening out of the granule. After a number of young rods had appeared, the drop was sucked up into a sterile capillary pipette containing a little sterile distilled water, and was expelled on to a slant of Petragnani's medium No growth appeared after incubation.

At this point, the writer decided to seek advice and criticism from Dr. Kahn, who kindly examined some of the stained preparations. He expressed a favorable opinion of the material shown and of the interpretations made, but said that the hanging-drops were by far too large to be considered as a vehicle for proof of a previously undemonstrated form. He suggested undertaking a single-cell study of the zoogleal forms, as the only adequate means of obtaining proof of their origin, and consented to have the writer use a single-cell apparatus in his laboratory at Cornell Medical College.

SINGLE-CELL STUDIES

1. Morphology of Diphtheroids

Dr. Kahn advised first practising the exacting technic with yeast and saprophytic diphtheroids. Since diphtheroids have so often intruded themselves into cultures of even the most earnest workers on morphological problems, they were considered well worth studying in themselves, especially as they possess many morphological features in common with the acid-fast bacteria. Sixteen strains which had been isolated from the air by Dr. John C. Torrey, were employed in this study. One of these, No. 7, was studied in especial detail.

In addition to single-cell preparations, a series of stained preparations was made daily in duplicate from three different media, and stained (1) by the Ziehl-Neelsen technic using Loeffler's methylene blue as counterstain, and (2) by the author's triple stain method. The media used were 0.75% hormone agar (a medium favorable to diphtheroids), 1% maltose veal infusion broth pH 6.0, and Corper's glycerol egg yolk medium(used in culturing tubercle bacilli).

Starting with a small group of young rods or short coccoid rod forms in microdroplets, it was noted that, after a few days' incubation, one or more of the group became refractile, accompanied by a thickening of the outer cell membrane. One or more tiny granule bodies appeared within these "globoid bodies" The globoid bodies stained a burgundy wine color with carbol fuchsin. Later, the small granule bodies appeared to have been extruded from the larger globoid forms, to have increased in number, and, in some instances, to have developed delicate hair-like extensions, which thickened into rods.

It would seem that with the diphtheroids, as with tubercle bacilli, new rods can develop from small granules, as well as by fission. Zoogleal forms were found in both stained and unstained preparations. The zoogleal forms usually contained granules, and, sometimes, a larger globoid body In a hanging-drop of normal guinea-pig serum, the outer membranes of some of the large globoid forms containing small granules, were observed breaking down into zoogleal forms In another drop, young rods were clearly seen growing from within such large bodies, apparently before the inner granules had been extruded The main morphological difference between the diphtheroid and tubercle organisms was in their rod forms. The diphtheroid rods were often plumper, prone to develop clubbed forms, and of a rigidly straight shape rather than slightly curved. Moreover, the group arrangement of the rods was different. The diphtheroid rods stained green rather than red or blue with the triple stain. In both diphtheroids and tubercle bacilli, the essential and permanent element of the organism appears to be a small but organized granule. That types, similar to the zoöglcal forms and globoid bodies described here, may be found in other organisms is indicated by the recent work of Dienes and Smith on *Bacteroides.*[8]

2. Zoogleal Forms of the Tubercle Bacillus

A single-cell study of zoogleal forms of the tubercle bacillus was then begun. An attempt was made to capture single zoogleal clumps within microdroplets, and to trace their development, following the technic of

Kahn. Accordingly, Long's medium pH 7.1, plus 10% horse serum, was chosen as a medium for the microdroplets. The medium was first subjected to centrifugation in a high speed centrifuge (4,000 R.P.M.) for one hour, in order to clarify it as much as possible, and free it from particles which might cause confusion in microscopic observations. A little of the top portion of the supernatant fluid was sucked up with a sterile capillary pipette and placed in a small, specially cleansed, sterile test tube. To this medium was added the suspension of organisms. The organisms had been pipetted from fluid beneath the pellicle of a 3½ weeks old culture of tubercle bacilli grown on Long's medium. Tests were made on all materials for the presence of contaminants by inoculating them into 0.75% hormone agar. The strains used were H37 (single-cell strain of Kahn), and PB 15, a virulent human strain obtained from the laboratory of Dr. Florence Seibert.

Zoogleal forms were isolated in a number of drops, but failed to develop beyond an occasional increase in the number of granule elements. Lung broth, which had previously stimulated regeneration of zoogleal forms into rods, was considered too difficult to clarify for use in single-cell work. Vitamin B complex (Merck) and guinea-pig serum were added to Long's medium for enrichment, but these failed to stimulate further development. Many months were consumed in trials and failures, using both pure cultures and tuberculous chest fluids. During this time, Dr. Nine Choucroun was helpful in giving advice and technical aid in preparing droplets and improving the writer's method of sealing the coverslips so that the preparations were less likely to dry up than previously.

A final control experimental series was then planned and followed through with pure cultures suspended in Eli Lilly's distilled water and incubated in small test tubes.

TECHNIC AND RESULTS OF CONTROL EXPERIMENT WITH SPECIAL DISTILLED WATER

The Effect of Pure Distilled Water (Eli Lilly) upon Four Strains of Human Tubercle Bacilli, H37 (Kahn), PB 15 (Seibert), Conti (Alexander-Jackson), and Cyron (Alexander-Jackson).

Experimental Procedure

A. To each of four sterile 100 ml. Erlenmeyer flasks containing glass beads, were added (1) 1 ml. special distilled water, and (2) a small colony of tubercle bacilli from Corper's medium. These ingredients were shaken well by hand for several minutes, and more water added to make an emulsion up to 8 mls.

B. 0.5 mls. of the above bacillus emulsion was pipetted into each of five tubes. There were five tubes for each strain. To every one of the twenty tubes was

added 2 mls. of distilled water. The rack of test tubes was then placed in the incubator at 37° C.

C. When pipetting the emulsions into the tubes, a drop from each emulsion was placed on 2 clean, flamed, previously unused slides. These drops were not rubbed with a loop, but allowed to dry in the air. One slide of each pair was stained by the usual Ziehl-Neelsen method, and the second one was stained by the triple stain technic. Stained preparations were also made of distilled water alone.

D. After 3 days of incubation, one tube only for each strain was opened, and two drops placed respectively on two slides, and stained as described above.
After 6 days, stained preparations were made from tubes *1* and *2*.
After 9 days, from tubes *1*, *2*, and *3*.
After 2 weeks, from tubes *1*, *2*, *3*, and *4*
After 4 weeks, from tubes *1*, *2*, *3*, *4*, and *5*.

E Subcultures were made on Corper's medium from each #5 tube of the series, to test the viability of the organisms.

F. Sterility tests were made from all the tubes on to hormone agar.

Results

1. Sterility Tests: Strain H37. . . . . .tubes 1 and 4, contaminated
PB 15. . . . . .no contamination
Conti . . . . .no contamination
Cyron . . . . .no contamination

2 Viability Tests: Strain H37. . . . . no growth
PB 15 . . . . good growth
Conti . . . . no growth
Cyron . . . . good growth

(cf. PLATE 1, FIGURE 4; PLATE 2, FIGURES 5–8)

3 Stained Preparations:

In the case of each one of the above strains, when stained by the triple stain method, zoogleal forms were seen on stained slides made on the 9th day of incubation, whereas none had been noted on the previous slides.

(cf. PLATE 1, FIGURE 3.)

Returning, once again, to an attempt to culture zoogleal forms in microdroplets, it occurred to the writer that, since very young animals and babies were particularly susceptible to tuberculosis, embryo juice added to Long's medium might be as stimulating for a reversion of zoogleal forms to rods as glycerol lung broth. Through the kindness of Dr. Lillian Baker of the Rockefeller Institute, some sterile chick embryo juice was obtained. Dr. Baker suggested that the medium be shifted slightly to the acid side as a probable additional aid to growth. The medium which did give the desired result consisted of Long's medium pH 6.8, 10% embryo juice, and 2% horse serum. The strain of tubercle bacilli studied with this new medium was PB 15.

A microdroplet containing a few good-sized zoogleal forms was placed under the oil immersion objective of a microscope with camera attach-

ment. Successive photomicrographs, taken of two of these, demonstrate that the zoogleal form of the tubercle bacillus can return to the rod form (PLATE 1, FIGURES 1–2). Thus, we have not merely a degeneration phenomenon (PLATES 3–6), which might be only of academic interest, but also, under suitable conditions, a regeneration into the rod form, a phenomenon which may prove significant from a medical and public health standpoint. It may be of interest to mention that a spinal fluid (tested for contamination) from a case of tuberculous meningitis, in which no acid-fast rods were found at first, was, at this time, swarming with zoogleal and granule forms. These were seen both in hanging-drops and in preparations stained with the triple stain.

DISCUSSION

Much additional work with both pure cultures and clinical material is necessary to understand more fully the part played by the zoogleal forms in tuberculous infection. We still do not know just what happens to the organisms when acid-fast bacilli disappear from sputum in arrested cases, only to reappear in large numbers with the return of active disease. Are most of them destroyed, or does a portion of them undergo a dissociative change in response to an environment temporarily unfavorable to the maintenance of the characteristic rod form of *Mycobacterium tuberculosis?* Another mystery is the difficulty, at times, of finding any acid-fast bacilli in stained preparations from pathological material, where one would expect to find many. Frequently, a definite diagnosis of tuberculosis is delayed by failure to find typical red-staining rods, until some time after the first laboratory examination.

In approaching these problems, the writer is only too well aware of the great difficulties which lie along the path, a path made thorny by the ever present danger of confusing the "real thing" with non-tuberculous elements—in particular, the diphtheroids. The utmost care, therefore, was taken to exclude their presence and to maintain a stoic skepticism until repeated close observations had been made. The work described above covers a ten-year period.

SUMMARY AND CONCLUSIONS

1. *Mycobacterium tuberculosis* exists, not only as rods or granules, but also as a zoogleal plasmodium consisting of granules or larger globoid bodies surrounded or enmeshed by amorphous material.

2. These zoogleal forms are not revealed by the usual stain technics (unless acid-fast). They are made observable by the new staining technic mentioned above.

3. Zoogleal forms have been observed repeatedly in unstained material as well as in preparations stained by the triple stain method.

4. Single-cell studies and electron microscope photographs of material from pure cultures, indicate that zoogleal forms, under suitable environmental conditions, are able to revert to rod forms and *vice versa*.

5. It has been observed that diphtheroids are also able to enter a zoogleal state, and to revert to rod forms.

6. The demonstration of forms which are not revealed by the usual technic may throw some light on the question of the apparent disappearance of acid-fast bacilli or their paucity during the course of some tuberculous infections.

I take pleasure in thanking Dr. Morton C. Kahn for his valuable advice and criticism, and for making available the single-cell apparatus; Dr. John C. Torrey and Dr. Ralph Nauss, for their helpful interest, and Dr. Nine Choucroun for encouragement and aid.

Most of the pathological material was obtained through the kind consent of Dr. J. Burns Amberson, head of the Tuberculosis Service at Bellevue Hospital; also some from the late Dr. Jesse G. M. Bullowa of New York University and Harlem Hospital, (Pneumonia Service), the Branch Laboratory of the New York State Department of Health, and Dr. James Edlin of Polyclinic Hospital.

The photomicrographs and motion pictures were taken by Mr. Jack Godrich, artist and photographer of the Department of Zoology, Columbia University, and the electron microscope photographs were taken by Dr. James Hillier of the R C.A. Laboratories, Princeton, N. J. To them I am most grateful.

BIBLIOGRAPHY

(Organized according to subject matter)

Morphology

1. **Alexander, E. G.**
1934. Developmental morphology of human and avian tubercle bacilli on Bordet-Gengou medium. Proc. Soc. Exper. Biol. and Med. **31**: 1103.

2. **Alexander-Jackson, E.**
1936. Studies on the dissociation of tubercle bacilli with special reference to the avian and human types. Amer. Rev. Tuberc., **33** (6): 767.

3. **Arloing, F., A. Dufourt & Marlatre**
1926. Études sur les variations morphologique et pathogenes du bacille de la tuberculose. Paris Medical, **16**: 22.

4. **Arloing, F., L. Thevenot, & J. Viellier**
1937. Influence de la decompression atmospherique et de l'anaerobiose sur les cultures liquides homogenes du bacille tuberculeux humain. Compt. rend. Soc. de Biol., **124**: 161.

5. **Arloing, M. S.**
1908. Variations morphologique du bacille de la tuberculose de l'Homme et des Mammiferes, obtenu artifiellement. Compt. rend. Acad. Sciences, **146**: 100.

6. **Claypole, E.**
1913. On the classification of the streptothrices, particularly in their relation to bacteria. Jour. Exper. Med., **17**: 99.

7. **Coppen-Jones, A.**
1895. Ueber die morphologie und systematische stellung des tuberkelpilzes, und über die kolbenbildung bei Aktinomycose und tuberculose. Centralblatt f. Bakt., **17**: 70.

8. **Dienes, L. & W. E. Smith**
1944. The significance of pleomorphism in bacteriodes strains. Jour. Bact., **48** (2): 125.

9. **Dreyer, G., & R. L. Vollum**
1931. Bacillaemia in experimental tuberculosis. Lancet, **220**: 1015.

10. **Foulerton, A. G. R.**
1910. The streptothrichoses and tuberculosis. Lancet, **1**: 552.
1912. As to the nature of the parasites of leprosy and tuberculosis. Brit. Med. Jour.: 300.

11. **Gulbrandsen, L. F.**
1935. Invasion of body tissues by orally ingested bacteria, and the defensive mechanism of the gastro-intestinal tract. Amer. Jour. Hyg. **22**: 257.

12. **Kahn, M. C.**
1929. A developmental cycle of the tubercle bacillus as revealed by single-cell studies. Amer. Rev. Tuberc. **20**: 150.

13. **Kahn, M. C., & J. F. Nonidez**
1936. The role of non-acid-fast rods and granules in the developmental cycle of the tubercle bacillus. Amer. Rev. Tuberc. **34**: 361.

14. **Karwacki, L.**
1931. Bacille tuberculeux comme forme evolutive d'un streptothrix. Centralblatt f. Bacteriol., orig. **119**: 369.

15. **Kredowsky, W. J.**
1937. Variabilité du groupe d'actinomycetes, et son rapport a la doctrine de la nature mycelienne de la virus tuberculose et de la lepre. Centralblatt f. Bakteriol. Ref. **124**: 352 (orig. 1936. Giorni. Batter. **17**: 289.)

16. **Lurie, M.**
1939. Studies on the mechanism of immunity in tuberculosis. Jour. Exper. Med. **69**: 576.

17. **Mallassez, L., & W. Vignal**
1883. Tuberculose Zoogloeique. Arch. de Physiol. **ii**: 369.

18. **Much, H.**
1907. Ueber die granulare nach Ziehl nicht farbare form des Tuberculosvirus. Beitr. z. klin. d. Tuberk. **8**: 85-99, 357-368.
1908. Granula und Splitter. Ibid. **2**: 70.

19. **Morton, H. E.**
1940. Cornybacterium diphtheriae—a correlation of recorded variations within the species. Jour. Bact., **4** (3): 177.

20. **Oerskov, G.**
1932. Eine morphologische Untersuchung über das Initialwachstum des Tuberkelbazillus. Z. f. Bakt., Par. & Inf. **123**: 271.

21. **Vera, H. D., & L. F. Rettger**
1940. Morphological variation of the tubercle bacillus, and certain recently isolated soil acid-fasts, with emphasis on filterability. Jour. Bact. **39** (6): 659.

22. **Wykoff, R. W. G.**
1934. Bacterial growth and multiplication as disclosed by micro motion pictures. Jour. Exper. Med. **59**: 381.

FILTERABILITY

23. **Calmette, A., & J. Valtis**
1930. Le virus tuberculeux (granulemie prebacillaire et bacillose). Ann. de l'Inst. Past. **44**: 629.

24. **Negre, H. L.**
1936. Research on the tuberculosis virus. Address given at the New York Academy of Medicine.

25. **Valtis, J., & A. Saenz**
1930. Sur la culture de l'ultravirus tuberculeux. Compt. rend. Soc. de Biol. **103**: 134.

26. **Vaudremer, A.**
1923. Formes filtrante des bacille tuberculeux. Compt. rend. Soc. de Biol. **89**.
Also of **Vera, H. D., & L. F. Rettger** cited under MORPHOLOGY.

DANGERS OF CONTAMINATION

27. **McCarter, J., & E. G. Hastings**
1939. The presence of avian tubercle bacilli in apparently pure cultures of diphtheroids. Jour. Inf. Dis. **64**: 297.

28. **Pinner, M.**
1933. Atypical acid-fast organisms. Jour. Bact. **25** (6): 576.

29. **Schwabacher, H.**
1933. Tuberculous bacillaemia (appendices). Med. Res. Council: 124. London.

STAINING

30. **Alexander, E. G.**
1932. Improved staining techniques for the demonstration of non-acid-fast tubercle bacilli and granules. Science **75**: 197.

31. **Alexander-Jackson, E.**
1944. A differential triple stain for demonstrating and studying non-acid-fast forms of the tubercle bacillus in sputum, tissue and body fluids. Science **99**: 307.

32. **Corper, H. J.**
1926. Methods of staining tubercle bacilli. Jour. Lab. and Clin. Med **11**: 503.

33. **Hollande, A. G. & G. Gremieux**
1928. La coloration vitale du bacille de Koch par le bleu de Nile. Compt. rend. Soc. de Biol. **98**: 1379.

34. **Kieffer, J.**
1921. New and easy method for demonstration of granules in tubercle bacilli. Amer. Rev. Tuberc. **5**: 662.

35. **Krylow, D. C.**
1912. Ueber die bedeutung und das vorkommen der Muchschen granula. Ztschr. f. Hyg. **70**: 130.
Much, H. cf. under MORPHOLOGY above.

36. **Yegian, D., & L. Baisden**
1942. Factors affecting the beading of the tubercle bacillus stained by the Ziehl-Neelsen technique. Jour. Bact. **44** (6): 667.

37. **Yegian, D., & K. R. Porter**
1944. Some artifacts encountered in stained preparations of tubercle bacilli. Jour. Bact. **48** (1): 83.

General Comment

38. **Barlaro, P. M.**
1929. Lecciones de Patologia Medica, Tratado de Tuberculosis. Editorial C.A.D.O.M., Distribuidor: Aniceto Lopez, Cordoba 2082. Buenos Aires (Refers to the work of Ferran and Fontes.)

39. **Hadley, P., E. Delves, & J. Klimek**
1931. The filterable forms of bacteria. Jour. Inf. Dis. **48**: 1.

40. **Koch, R.**
1882. The etiology of tuberculosis. Berliner, Klinisch, Wochenschrift **19**: 221.
1938. Reprinted in Medical Classics **2** (8): 821, 853.

41. **Manwaring, W. H.**
1932. Trend of medical bacteriology. Science **76**: 41.

42. **Rich, A. R.**
1944. The pathogenesis of tuberculosis. Chas. Thomas, Springfield, Ill

43. **Winslow, C-E. A.**
1932. The changing bacteria. Science **75**: 121.

PLATES 1-6

PLATE 1

Photomicrographs of the development of zoogleal forms into granules and rods; strain PB 15 (virulent human type), obtained from the fluid portion beneath the pellicle of a 3½ weeks' old culture grown on Long's medium.

Figure 1

a. Original zoogleal form in microdroplet.
b. Same form after one week.

Figure 2

a. Original zoogleal form in microdroplet.
b. Same form after one week.
c Same after 2½ weeks.
d. Same after 3½ weeks.
e Same after 5 weeks
(Long's medium was used in the microdroplets, modified by embryo juice)

Figure 3

a. Stained preparation (triple stain technic) of non-acid-fast zoogleal form in tuberculous chest fluid.
b. Stained non-acid-fast zoogleal form from a pure culture of *Mycobacterium tuberculosis* (virulent human strain H3 S). Note globoid bodies.
c. Semi-acid-fast young rods from tuberculous pleural fluid.

Figure 4

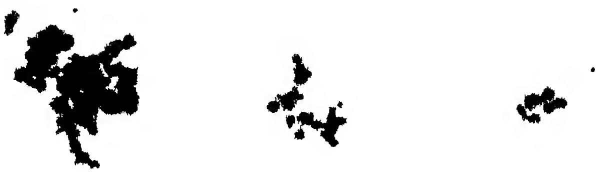
a. Intermediate form, H37 strain, in microdroplet
b. Intermediate form, PB 15 strain, in microdroplet.

a *b*

FIGURE 1

a *b* *c* *d* *e*

FIGURE 2



a *b* *c*

FIGURE 3

a *b*

FIGURE 4

FIGURE 9

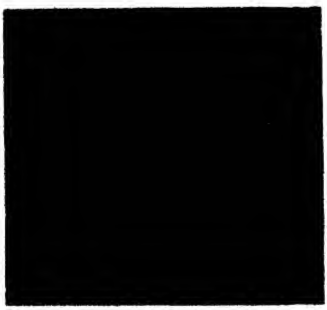


FIGURE 6

FIGURE 8

FIGURE 5

FIGURE 7

ALEXANDER-JACKSON MYCOBACTERIUM TUBERCULOSIS

PLATE 2

FIGURE 5. Microdroplet under low-power objective. Note organisms on the left. Magnification 450 ×.

FIGURE 6. Zoogleal form near the edge of the microdroplet. Strain PB 15. 1/16" oil immersion objective. Magnification 1250 ×.

FIGURE 7. Large zoogleal form in hanging drop. Strain H 37. Magnification 1000 ×.

FIGURE 8. Zoogleal and rod forms of a diphtheroid in a hanging drop of normal guinea-pig serum. Magnification 1000 ×.

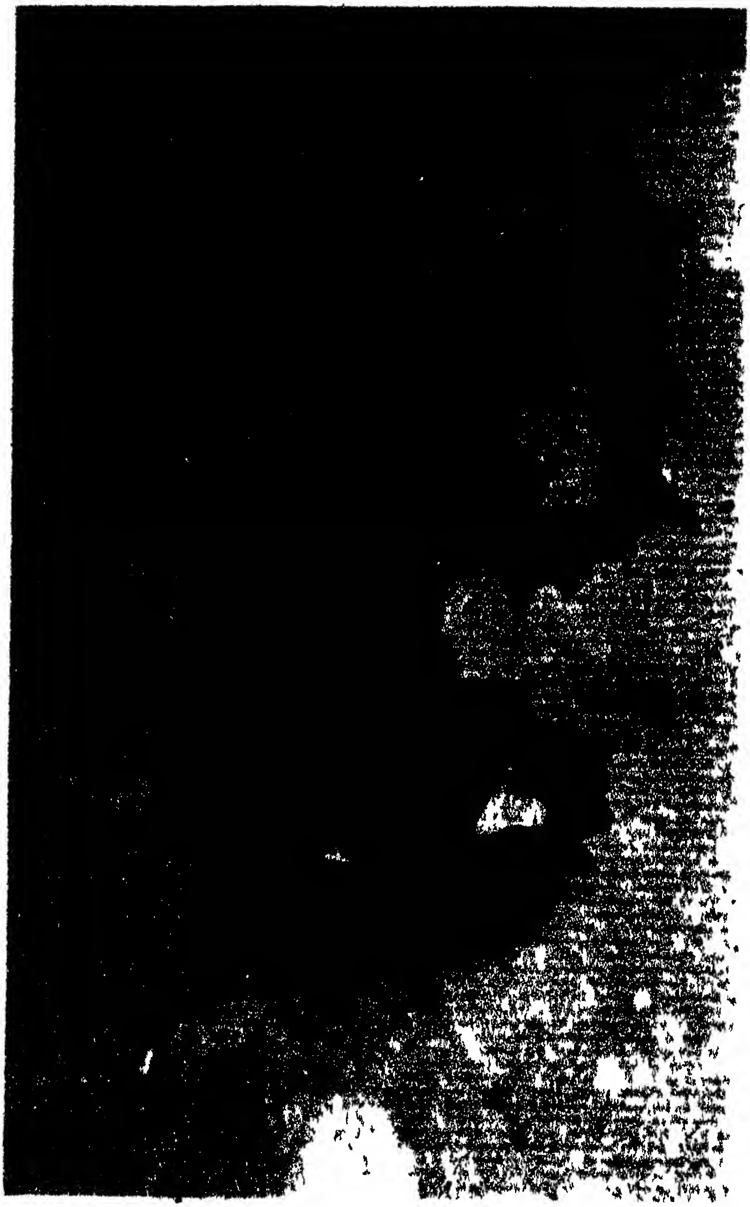
FIGURE 9. Photomicrograph of a stage micrometer under 1/12" oil immersion objective. Each small division represents 10 microns. Magnification 1000 ×.

PLATE 3

Electron microscope photograph

21,000 × enlargement of organisms from the fluid portion of a 3½ weeks' old culture grown on Long's medium Note "ghost" bacilli with amorphous material, some of which appears to be escaping from the rods

ALEXANDER-JACKSON MYCOBACTERIUM TUBERCULOSIS



ALEXANDER-JACKSON MYCOBACTERIUM TUBERCULOSIS

PLATE 4

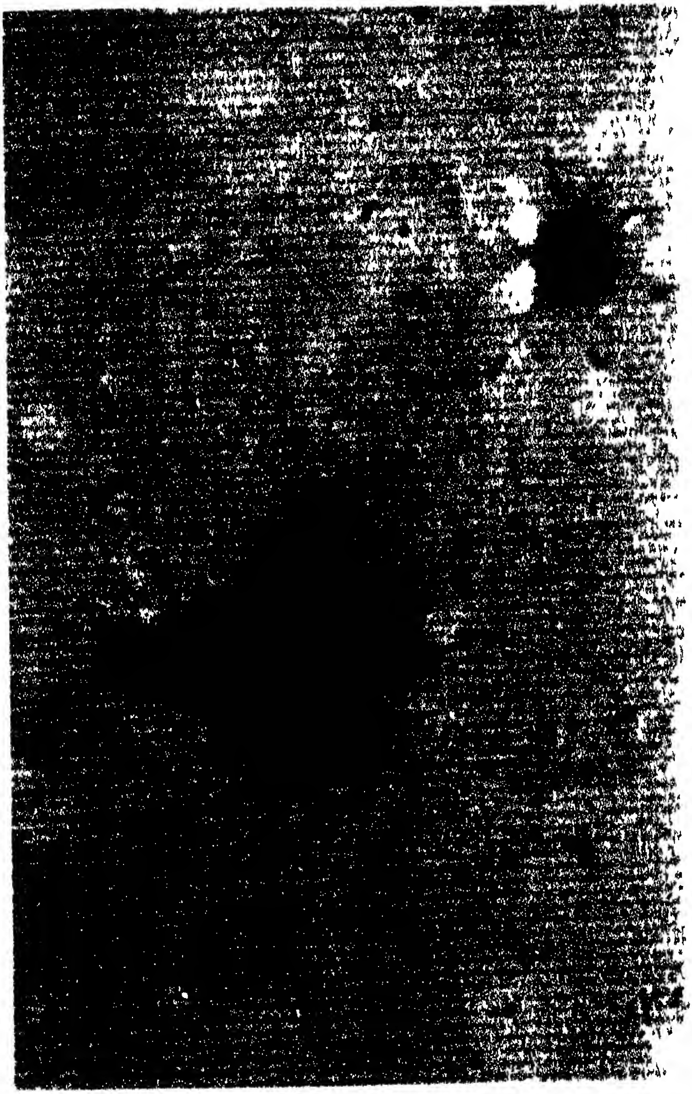
Electron microscope photograph

21,000 × enlargement of material from the same culture shown in Plate 3 Note rod forms degenerating into an amorphous mass.

PLATE 5

Electron microscope photograph

28,400 × enlargement from the culture shown in Plates 3 and 4 Note slender rod with escaping inner protoplasm



ALEXANDER JACKSON MYCOBACTERIUM TUBERCULOSIS

ALEXANDER JACKSON MYCOBACTERIUM TUBERCULOSIS

PLATE 6

Electron microscope photograph

28,400 × enlargement from the culture shown in Plates 3–5. Note three round dark bodies in small amorphous clump at the left.

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES

VOLUME XLVI, ART. 3. PAGES 153–184

JUNE 30, 1945

THE EFFECT OF ACTIVITY ON THE LATENT PERIOD OF MUSCULAR CONTRACTION*

BY

ALEXANDER SANDOW†

CONTENTS

* Awarded an A. Cressy Morrison Prize in Natural Science in 1944 by the New York Academy of Sciences. Publication made possible through a grant from the income of the Ralph Winfred Tower Memorial Fund.
The research was supported in part by a grant from the Penrose Fund of the American Philosophical Society.
† Department of Biology, Washington Square College of Arts and Science, New York University, New York City.

INTRODUCTION*

This research deals with the effects of muscular activity on the various features of that part of the contractile response known as the mechanical latent period—the time interval between the instant of application of the stimulus to the muscle fibers and the instant at which the first sign of their tension development appears. Since a burst of contractile activity of a muscle not only alters the latency behavior in subsequent contractions, but also causes marked and quite well-known changes in the muscle's internal chemical milieu, the possibility arises of attempting to elucidate the nature of the latent period in the light of the correlation between the activity-induced latency and chemical modifications.

The latent period is not, as has been generally thought, a time of complete mechanical quiescence. Actually, as first demonstrated by Rauh (1922), a frog muscle relaxes very slightly during the latter part of its latent period, just prior to the development of tension. This "latency relaxation" (abbreviated LR) involves a negative tension change in frog sartorii of, at most, about 20 mg., i.e., only about 0.05% of the positive tension output at the peak of a maximal twitch. This is so minute that even the most sensitive mechano-optical registering levers, such as, for example, Rauh used, can hardly do more than detect the change. It is obvious that any thorough-going investigation of the latent period must include studies of the LR, and that the successful development of such studies requires a new method for registering this phenomenon that will accomplish much more than merely record its presence.

In an earlier paper (Sandow, 1944b; hereafter to be designated "paper I"), I have discussed certain introductory aspects of the problem of the latent period which center around the LR. In this paper, there is described in great detail the new piezoelectric, cathode-ray oscillographic technique that has been devised for recording the LR. A brief summary of this method will be given here. A frog sartorius muscle, under slight initial tension, is connected to the stylus of a piezoelectric unit which is actually the cartridge of a crystal phonograph pickup The stylus (equivalent, in phonographic usage, to the "needle") is mounted in the chuck of the pickup, so that, from the standpoint of mechanical myography, it acts like an isometric lever. When the

* Preliminary reports of some of the results of this paper appear in Sandow (1942a, 1942b, 1943)
The author gratefully acknowledges the technical assistance of Mr. A. G. Karcsmar in completing the experimental work, as presented here

muscle is stimulated, the resultant tension changes (first, negative, during the LR; and then, positive, as contraction develops) are converted by the pickup into a corresponding electric pulse which, after suitable amplification, actuates a cathode-ray oscillograph. Thus, the latency mechanical variations appear as visible deflections which are traced out during single sweeps on the cathode-ray screen. The recording system acts, in effect, as an electronic lever with an extremely high magnification, for, with the usually employed voltage amplification of the amplifiers of some 5,000 fold, the length changes that accompany the tension alterations of the muscle during the LR are converted into approximately 500,000 x enlarged oscillographic deflections Thus, a typically obtained LR deflection of, say, 3 cm., corresponds to an increase in muscle length of 0.06 μ. The sensitivity and reliability of the apparatus are such that LR changes in the muscle of the order of 0.002 μ in length, or 0 07 mg. in tension, can be easily determined. The latency mechanical changes, which, in general, run their course in 3–4 ms., are spread out on a horizontal baseline of about 7 cm length, and a simultaneously impressed timing wave of 10,000 cycles/sec. permits latency time intervals to be measured with a precision of $\pm$ 0.02 ms.

PLATE 1 reproduces a photograph of a typical record to which has been added the symbols of the measured variables. The latency record begins at the extreme left, coincident with the instant of stimulation. After a quiescent period measured by L_R, the latency relaxation occurs, reaching a final depth measured by R, and is then reversed as contraction sets in. It will be noted that the original record also includes the intense band of light at the upper right which is an optical lever indication of the initial tension of the muscle, and, the much less intense, broader band of light below, which measures the peak isometric tension developed in the muscle's contraction. Thus, each record involves six latency measurements (four time intervals: L_R, L_0, L, and L_1; and two LR magnitudes, R_0, and R) and two tension measurements (initial tension, and peak developed tension, T).

These variables have been studied in frog sartorii as a function of the strength of the shock and of the initial muscle tension (paper I) and the following will summarize the experimental generalizations of this research concerning latent period mechanisms that must be known for an understanding of the present contribution dealing with the effects of activity. (1) The latent period includes three more or less distinct processes before the actual tension-process of contraction begins: (a) the LR-induction process which occurs during the interval L_R; (b) the LR process, and (c) the tension-induction process which may be

identical with the LR process itself. (2) Tension development is already under way during the latter part of the LR, and L_0 represents the earliest instant at which the onset of tension development can be detected; thus, L_0, L, and L_1 all represent different mechanical latencies, depending on the instant chosen to mark the end of mechanical latency. At any rate, they give a sort of chronology of the very earliest moments of tension development and, in this respect, it will be convenient to refer to them as a group, with L_T symbolizing the set. (3) The LR is an externally evident mechanical sign of some function of myosin, the contractile protein of muscle. It will be seen that the research to be presented here leads to results that give further substantiation to the above conclusions. Furthermore, the present work gives support to a previously made inference that the LR represents an enzymatic intermediary composed of myosin and its substrate, adenosinetriphosphate (ATP), during the existence of which the myosin is being energized for contraction; for, as will be discussed in detail later, the effects of activity on certain aspects of the LR are interpretable as being due to corresponding changes in the rate of hydrolysis of ATP brought about by the activity-induced modifications of the muscle's internal chemical milieu.

METHOD

The general electronic apparatus for recording the muscle's mechanical behavior has been outlined above. Certain physiological details of the procedure will now be given. The sartorius muscle of medium sized *Rana pipiens* was used in all experiments. After removal from the body, the muscle was equilibrated for at least one hour in several changes of oxygenated Ringer's solution buffered to about pH 7.1 with phosphates (12.4 mg P %). It was then set up in a moist chamber, with its attached pubic symphysis clamped in position near the bottom of the chamber and its tibial end extended upward and connected by a fine metal chain to the externally placed crystal pickup. The muscle was kept at a constant resting tension of about 3 gms., this being optimum for developed tension output resulting from stimulation. Stimuli were thyratron-tripped condensor discharges with a time constant of about 0.1 ms. and they were obtained from a stimulator which could be adjusted to give single shocks, or tetani of variable frequency and duration. Generally, maximal shocks directly stimulating the muscle fibers were used and were applied through Ag-Ag Cl electrodes placed so that the anode was at the muscle's pelvic end and the cathode, some 30 mm. distant, near the tibial end. The muscle was kept at a con-

stant temperature, generally about 22.0° C., by immersion of its chamber in a water bath held constant to within 0.01° C.

Several types of activity and recovery sequences were used in this research. They will be described individually in connection with the relevant experiments discussed below. But, in general, any one experiment involved three periods: (1) A pre-activity period during which the muscle was subjected to two maximal twitches, separated by an interval of four minutes. These responses served to obtain records of the muscle's normal rested behavior. Great care was taken to avoid any stimulation of the muscle prior to these tests, and if the muscle had shown signs of activity, it was allowed to rest for a time—never less than fifteen minutes—commensurate with the amount of activity, before the pre-activity responses were recorded. (2) Four minutes after the second pre-activity twitch the activity period was begun, and continued for the time required to run through the series of one or more rather closely spaced maximal twitches or tetani making up the particular activity. Since the very first latency response of the activity period was obtained from an essentially rested muscle, its measurements were averaged together with those of the two pre-activity twitches to obtain data of the mean rested behavior. (3) Except for certain types of experiments discussed later, the activity period was followed by a recovery period, during which the muscle was subjected to a series of maximal twitches separated, except at the very beginning, by relatively long time intervals and which served to determine the immediate effect of the activity and the subsequent course of recovery therefrom. Since the test for the latency behavior, at any instant, necessarily involves an entire twitch response, it is obvious that it was impossible to prevent some activity occurring during the recovery period. But, as control tests proved, the amount of activity necessarily needed to test for the course of recovery was so small and so spread in time in comparison with that of the activity period that its interference with the progress of recovery could be neglected.

RESULTS

Effect of a Single Tetanus

The typical effect of a tetanus of medium duration is shown in FIGURE 1. In this particular experiment, a 3 sec. tetanus having a frequency of 45 shocks/sec. was used and it was applied at the instant corresponding to the zero of the time axis. The points plotted at this instant are the averages obtained from the pre-activity responses and thus characterize the muscle at the moment of application of the

tetanus. The remaining points represent the behavior of the muscle in test twitches at the following instants after the beginning of the tetanus: 15 and 30 sec., 1, 3, 5, 10, 20, 30, and 45 min., and thus indi-

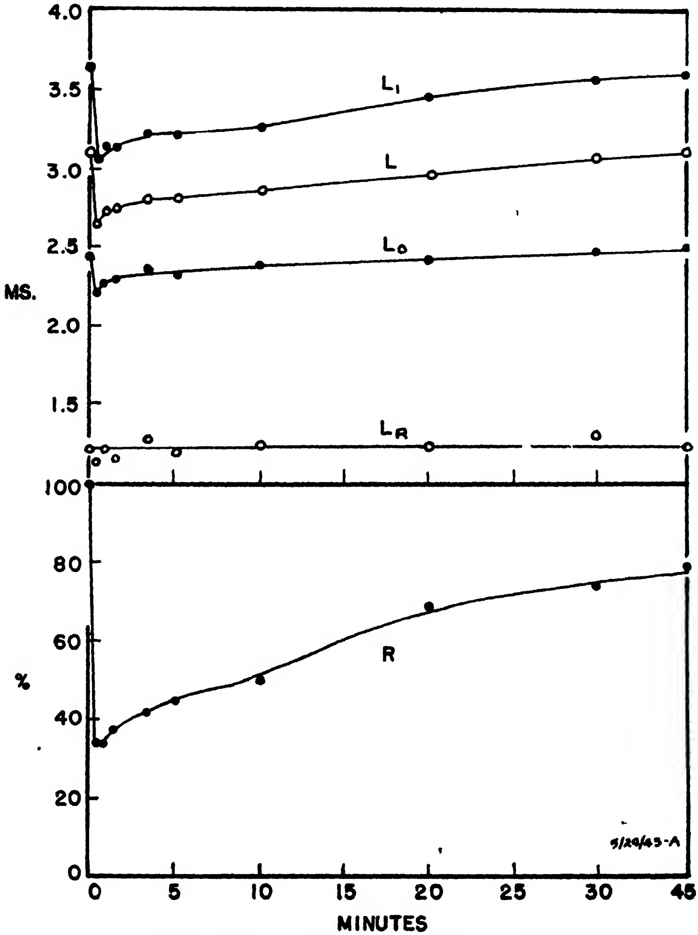


FIGURE 1 Effect of a 3 sec tetanus (45 shocks/sec.) on the latency variables The tetanus was applied at the zero of time Muscle at a constant initial tension of 3 gms. Temperature, 22° C. Note change in scaling of the time axis toward the end of the interval plotted.

cate the changes immediately due to the tetanus and subsequently occurring during the recovery period. PLATE 2 reproduces the strip of film of the original series of records which includes the photographs

made at the beginning of the tetanus (left frame) and at the 15 sec. recovery point (right frame). Observation during the actual recording shows that the sweep of the "tetanus" photograph having a clear L R deflection is the latent period response of the muscle to the very first shock of the tetanus. (The other irregular sweeps are traced out during later phases of the tetanus and thus have no meaning for the present analysis.) Since, as previously mentioned, this first response of the tetanus is that of a rested muscle, it may be used as a control for comparison with the latency response of the next photograph taken just after the termination of the tetanus in order to determine the immediate effect of the tetanus. Even without measurement, it is apparent that L and especially R have been markedly decreased by the tetanus; and actual measurement of these records may be used to verify some of the other observations that will be given detailed discussion below.

FIGURE 1 demonstrates that the tetanus has a pronounced and regular effect on all the variables studied except L_R. The study of the results of very many experiments similar to the present one proves that the fluctuations in L_R, such as appear in FIGURE 1, are purely random. Indeed, not infrequently, the measured values of L_R associated with an activity test are constant within the precision of reading the records (about $\pm$ 0.02 ms.). Thus, we must conclude that activity of the order of that used in this experiment has no effect on L_R.

Quite different is the behavior of the time intervals of the L_T set, L_0, L, and L_1. All of these are decreased by the tetanus and tend during the recovery period to regain their pre-activity values. The tension variables, R_0 (not plotted), R, and T, also are modified, both R's sharply decrease immediately after the tetanus and then during the recovery period they increase back toward their pre-activity values; T (not plotted) exhibits an initial increase followed by a gradual decrease. It may also be noted that the ratio R_0/R remains practically constant at 55% throughout the experiment.

Although the changes in the various L_T's are, at most, of the order of only a few 0.1 ms., there is no doubt as to their reality, for the immediate decreases in the L_T's followed by a more-or-less complete and gradual restoration of their pre-activity values during the recovery period have been obtained, without exception, in every comparable experiment. Furthermore, changes such as those observed would be expected to be statistically significant within even a single experiment, since, as demonstrated in paper I, the statistical error of any single latency time interval is about $\pm$ 0.02 ms. Similar arguments establish the validity of the recorded variations of the R's, despite the fact that

they also are very minute, involving, at most, an absolute change of decrease in tension of about 15 mg., or an equivalent change of increase in length of about 0.05 μ.

FIGURE 1 shows that the process of restoration of the L_T's and R to their pre-activity values during the recovery period follows a sigmoid curve that includes a sort of shoulder between the 3rd and 10th min. points. Despite the high order of precision of the values measured in any single response, it has so far been found impossible to decide whether this shoulder, at least as presented in FIGURE 1, is an invariable part of the recovery process, since different muscles vary greatly in respect to their recovery changes. In some experiments, the shoulder is completely absent, the recovery curves thus appearing logarithmic instead of sigmoid; in others, the shoulder, although beginning at about the same moment as indicated in FIGURE 1, extends over a much shorter time interval. It seems that, in general, the shoulder exists, but that it varies a great deal in definiteness and in duration.

Effect of a Series of Tetani

In this type of experiment, the muscle was subjected to a series of some 15 or more maximal tetani spaced at intervals of about 7 sec. Each tetanus stimulus was 2 sec. long and had a frequency of 45 shocks/sec. There was no recovery period after the activity, for the interest in this experiment is in the behavior of the muscle within the approximately 90 sec. period of seriate tetanus responses. (This technique was used in order to obtain latency data to compare with the pH changes demonstrated by Dubuisson (1937) in similarly activated frog sartorii, and about which more will be said later.) Records of the muscle's behavior were made by photographing the very beginning of each tetanus response, as was done in obtaining the left frame of FIGURE 1. Thus, the clear latency response of each record, representing the state of the muscle in the first twitch of the corresponding tetanus, indicates the accumulated effects of all of the preceding activity.

The results of a typical experiment are presented in FIGURE 2. It will be noticed, first, that L_R, unlike the other variables, does not change in any regular manner. The fluctuations are relatively pronounced late in the series of tetani where R is getting smaller, as is to be expected in view of the fact that the value of L_R is difficult to measure when the LR is not deep and thus begins very gradually. Nevertheless, the results of many seriate tetanus experiments indicate, as in the experiment dealing with a single tetanus, that the fluctuations in L_R are random, and we, therefore, conclude that L_R is not affected by the set of

tetani. This conclusion does not hold, however, if the tetanus series is very long and the muscle is very fatigued. Under these conditions, L_R seems to increase.

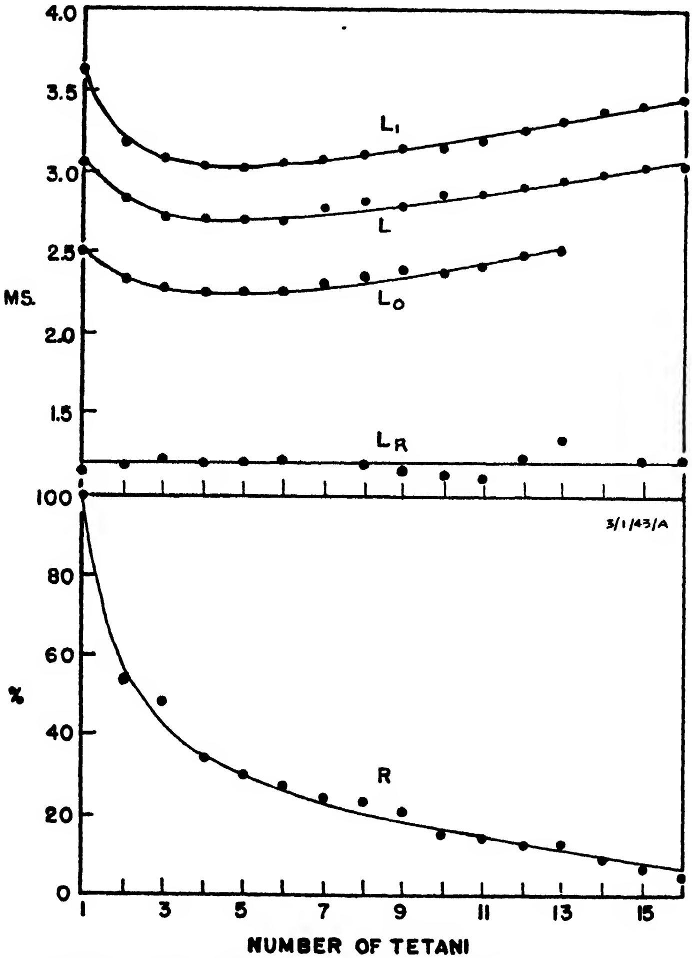


FIGURE 2. Effect of a series of 2 sec tetani (45 shocks/sec), spaced at roughly 7 sec intervals on the indicated variables. Initial tension, 3 gms Temperature 21 5° C

L_0, L, and L_1, all vary in parallel, as in the previous experiment, and this is further experimental justification for including L_0, despite its

occurrence within the LR, in the set of tension latencies, L_T. The tension latencies first decrease in response to the first four tetani, but, as the series of tetani continues, they then rise, at a smaller rate, however, than they had decreased. The initial decreases in the L_T's conform with the results of activity in the previous experiment. But the subsequent increases present a new aspect of special interest, for these increases are caused by continued activity, whereas in the experiment involving a single tetanus the increases of the L_T variables were caused by recovery from earlier activity.

R falls throughout the entire series of tetani This behavior is noteworthy for it demonstrates, in contradiction to the single tetanus experiment, that changes in R do not necessarily parallel those of the L_T's; for, in the present experiment, decreasing values of R after the fifth tetanus are associated with increasing values of the L_T's, while, in the single-tetanus experiment during recovery, R increases along with the L_T's. The value of R_0/R fluctuated very irregularly in this experiment, averaging, however, 69%. Other experiments show smaller random variations, about some average value, generally about 60–70%, and we thus conclude that this ratio remains constant in any one muscle during the entire activity series, although it may vary from muscle to muscle. Finally, it may be noted that the developed tetanus tension exhibits a gradual decrease with increase in the number of tetani.

The relation of these results to the pH changes of Dubuisson, previously mentioned, will be discussed later.

The Effect of a Series of Twitches

The experiments utilizing tetanus activity are limited, in that they do not yield any information concerning the change in behavior of the muscle during the tetanus, for the technique described above necessitates completion of the tetanus before introduction of the test twitch. Now the experiments of Dubuisson already mentioned involved continuous recording of the pH changes of the muscle both during, as well as for some time after, each tetanus. Furthermore, Lipmann and Meyerhof (1930) performed experiments concerned with pH changes in which the activity consisted of a 3 min. series of twitches at the rate of 20/min. It was, therefore, desirable to repeat the Lipmann and Meyerhof activity procedure here, not only to obtain a basis for comparing our results with theirs, but also because their activity schedule, when used in our experiments, permitted the registering of the latency behavior in each twitch and thus afforded a kind of continuous record of the changes during a burst of activity somewhat comparable to the

Dubuisson continuous pH recording during the tetanus. It is admitted that stimulation with a series of twitches at 3 sec. intervals is far different from stimulation with a tetanus whose frequency is 45/sec. But technical limitations of our method do not allow obtaining individual latency records separated by intervals less than 3 sec. Indeed, recording at this rate could only be accomplished by placing some 5 or 6 individual latency responses on a single photographic frame, avoiding overlapping by spacing the separate records at progressively lower levels of the cathode-ray screen through proper manipulation of the oscillograph's vertical positioning control between the successive twitches (PLATE 3).

The results of an experiment involving a series of twitches at the rate of 20/min. for 3 min. is given in FIGURE 3. To simplify the graphs, only those values of the activity period are given that are needed for adequate delineation of the general behavior during this period. How-

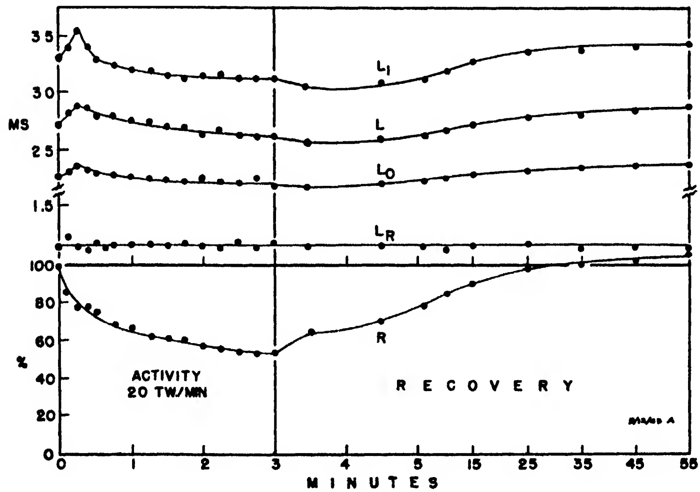


FIGURE 3. Effect of a 3 min series of twitches at the rate of 20/min Initial tension, 3 gms Temperature, 21 5° C. Note change in scale of time axis after the 5 min point

ever, all the points of the recovery period are plotted. Comparison of the results of FIGURE 3 with those of FIGURE 1 shows that the immediate net effects of the total activity in both cases are essentially the same. L_R remains constant, in fact, it does not change appreciably throughout the entire experiment; the members of the L_T set are all decreased; and R also is decreased. During the recovery periods of the two experi-

ments, the results are again similar except that the L_T's continue to decrease slightly but definitely for 1½ min. after the cessation of the activity twitches before commencing their characteristic recovery period increases. By far the most striking feature of FIGURE 3 is the fact that each of the L_T's increases slightly during the first 15 sec. of the activity period before they begin to decrease during the remainder of the activity. This effect is evident in the actual records of PLATE 3. Thus, although a moderate amount of activity—such as a 2 or 3 sec. tet., or the entire set of twitches of the current experiment—causes a net decrease in the values of the L_T's, a slight amount of activity—such as 5 or 6 twitches at 3 sec. intervals—causes an increase in these values.

The initial rises in the L_T's, and also the continued fall of these variables for a short time after the termination of the twitch activity series, involve changes generally of about 0.1 ms. and never more than 0.2 ms. Since the statistical error of the individual L_T values is about $\pm$ 0 02 ms., these changes are large enough to be statistically significant. Be that as it may, these features of the L_T's have appeared in all but one of the 20 twitch activity series experiments that have been performed. We must, therefore, conclude that they are characteristic effects of this type of experiment.

Localization of the Source of the Latency Response Changes

In considering this question, it is necessary first to eliminate the recording apparatus as the source of the changes. This is especially pertinent in respect to the possibility that the reductions in R may be a consequence of a reduction in amplification of the recording apparatus, due to the blocking effect of the intense electric pulses piezo-electrically developed, in response to the full developed tension of the twitches or tetani comprising the activity. This kind of artifact, however, may be definitely eliminated. For characteristic activity effects are registered by the amplifier even when it is kept completely quiescent during the activity period by grounding the activity output of the pickup. Furthermore, if, e.g., a tetanus pulse, is not grounded and some blocking does occur, independent tests show that, by the time the first post-activity test response is recorded, the amplification is again at its pre-activity value. Thus, the observed changes are affected by alterations in the state of the muscle and not of the recording apparatus.

Turning now to the muscular changes, it should first be recalled that, as shown in paper I, the LR is a function of myosin. The question now to be settled is: What part of the muscle structure is so altered by activity as to modify ultimately the LR behavior of the myosin? Since

the muscle is a complex organization involving not only a contractile system made up of myosin and its associated chemical energy reservoirs, but also various excitatory systems, and since, in general, any of these components may be locally or extensively affected by activity, it is clear that, *a priori*, there are several possible mechanisms whose changes might be finally reflected in the way the contractile machine responds. The evidence about to be presented, however, indicates that a general change due to activity in the contractile system itself is the source of the recorded latency alterations.

It is noteworthy that all the characteristic activity and associated recovery effects are obtained regardless of the means by which the muscle fibers are finally excited. Thus, direct stimulation of either normal or curarized muscles, and indirect stimulation of normal muscles with the electrodes, either on the muscle or on its nerve, all give essentially similar results. The absence of any special activity effects that might be correlated to the particular mode of stimulation, indicates that the general changes observed are rooted in the common mechanisms ultimately set in action by all these modes of stimulation— i.e., the action potential and contractile mechanisms of the muscle fibers. The possibility that changes in the action potential mechanisms are responsible for the latency changes is ruled out by the fact—to be treated in detail elsewhere—that the effect of activity on the electrical response follows a quite different pattern from that observed in the latency mechanical changes.

The fact that typical effects are obtained when, as in stimulation by way of nerve, the electrodes do not touch the muscle, suggests that the changes in the contractile system are extensive, rather than local alterations near the electrodes, even in those cases where the electrodes do rest on the muscle. This point is proved by two series of experiments, in the first of which the electrode polarity and position on the muscle used for the recovery test twitches was different from that used in the activity tetanus, and in the second of which the activity stimuli were applied either to the nerve or the muscle, while the recovery stimuli were applied, respectively, to the muscle or the nerve. In all experiments of these two series, the effects of activity were essentially the same, irrespective of whatever change in electrode arrangement was made in passing from the activity to the recovery stimuli.

Hence, the effects of activity cannot be attributed to modifications in the nature of excitation due to changes localized in the neighborhood of the electrodes used in the activity period, for if this were true, application of the post-activity test stimuli with a different spatial or elec-

trical configuration from that used during the activity period, would result in some variation or even a disappearance of the activity effects. Since the propagated disturbances evoked by the activity and post-activity stimuli, no matter how applied, would always cause to respond the entire length of each excited fiber, and since, as previously indicated, action potential mechanisms are ruled out, we must conclude that the latency changes following activity are attributable to alterations more or less evenly distributed throughout the whole contractile system of each fiber. The present analysis does not permit us to determine whether it is the myosin or the chemical sources of its energy that are modified by the activity, but we will return to this point in a later section of the paper.

Correlation of the L_T with the pH Changes Caused by Activity

Although the present research does not include any pH measurements made in parallel with the latency records, it is, nevertheless, possible to demonstrate that a definite correlation exists between the time course of the changes of our L_T's due to activity, and that of the pH of the muscle made by other workers in comparable experiments. All experimenters [Lipmann and Meyerhof, (1930); Meyerhof, Möhle, and Schulz, (1932); Dubuisson, (1937), (1939); Hill, (1940); Millikan, (1942)] agree that, in a muscle initially at about pH 7.0, the immediate net effect of a moderate burst of activity, comparable to our 3 sec. tetanus or 3 min. series of twitches, is an alkalinization of the muscle that is quite slowly reversed during subsequent recovery. The methods used for determining this pH sequence differ among themselves and hence the details of the sequence vary. Furthermore, in no case, does the particular activity procedure in all features correspond to ours. Thus, it is impossible to compare any of these results exactly to ours. The work of Hill (1940), however, includes an experiment in which a muscle at 0° C. and pH 6.5 was subjected to a 0.2 min. tetanus and the pH was followed manometrically by measuring the change in the CO_2 content in the gas space about the muscle. The results show that the muscle reached an alkaline maximum very soon after the tetanus and that this was gradually reversed during the next 50 min. This sequence of pH changes is, at least qualitatively, quite like the changes in the L_T's following a short tetanus, for here an immediate reduction in the L_T's is slowly reversed during recovery. In view of these comparisons, it thus appears that the greater the pH, the smaller are the L_T's.

Fortunately, a much more convincing basis for this correlation is afforded by the work of Dubuisson. He followed the pH changes of

an activated frog sartorius muscle by placing a glass electrode in contact with the film of Ringer's solution adhering to the muscle surface. This method has the advantage of a much smaller lag in recording than that present in manometric methods, but, due to unsteadiness in the base line, it cannot be used to follow the reaction changes during the whole of a prolonged recovery period. It is very well adapted, however, for registering the continuous course of the pH changes within a single tetanus or accompanying a series of closely spaced tetani, and such experiments will now be discussed in relation to the generally similar ones in the present research.

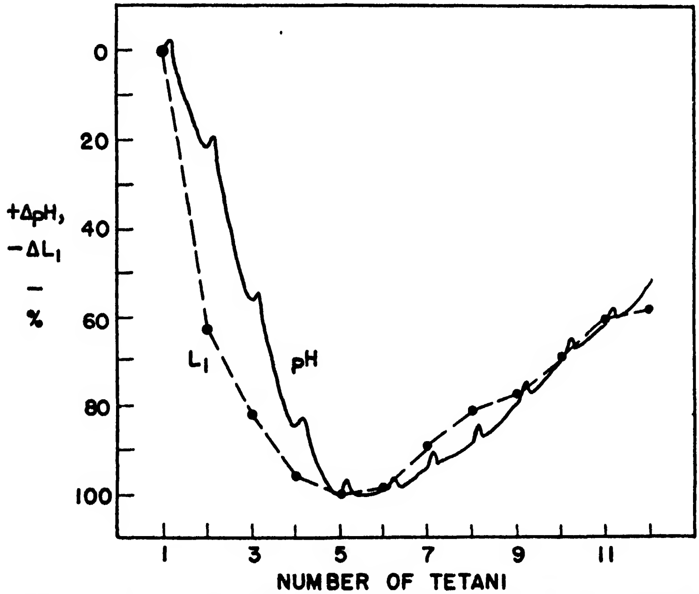


FIGURE 4. Comparison of the average behavior of L_1 in 12 experiments each involving a series of tetani (as in the experiment of figure 5), with the change in pH of a sartorius muscle (Dubuisson, '37) in a similar tetanus series. See text for further details.

FIGURE 4 presents the course of the pH changes at the surface of a muscle activated in a series of 2 sec. tetani at about 10 sec. intervals (Dubuisson, 1937; see also Dubuisson, 1939). The original results have been replotted here on a percentage basis with the maximum alkalinization taken as standard. Since only a single experiment is involved it is presumed that the results given by Dubuisson correspond to the typical behavior of the muscle, and may thus be compared with

the typical L_T changes of a similarly tetanized muscle graphed in FIGURE 2. Disregarding, for the present, the small reversible acidifications on Dubuisson's curve, it is evident that the general overall pH change shows many similarities to the L_T variations. These similarities are more definitely indicated by comparison with the second curve of FIGURE 4 which summarizes the average behavior of L_1, used here to typify the L_T set, in identical experiments on 12 different muscles. As in the pH curve, the points are plotted on a percentage basis, with the largest mean decrease in L_1 taken as 100%. There are several striking similarities in the two curves. Each exhibits a rather sharp drop up to the fifth tetanus and then a relatively slow rise. It must be emphasized that the coincidence at the fifth tetanus of the maximum change in L_1 fall and pH rise is not an arbitrary result of scaling (all that is arbitrary here is that each maximum change is set at 100%), but that this coincidence signifies a true parallelism in the course followed by the tension latency and pH changes.

It is of interest that the pH changes lag behind those of L_1 and that the lag is particularly marked at the start of the series of tetani. If the pH and tension latency changes are correlated in the sense of being different expressions of some common underlying alteration of the muscle, such a lag would be expected to occur. Previous discussion has indicated that activity causes changes in the L_T's by directly affecting the state of the contractile system. Since the tension latencies are an expression of the contractile response, it is clear they would reflect any activity changes practically instantaneously. The pH changes, as measured by Dubuisson, however, would necessarily show a lag relative to the fundamental chemical alterations that underly them. For the pH measurements are made in the film of Ringer's solution on the outside of the muscle, while the chemical changes, principally creatine-phosphate hydrolysis with an accompanying increase in pH and lactic acid formation with a decrease in pH, occur within the muscle fibers. Thus, the time required for equilibration of the outside and inside of the fibers would cause the pH measurements to run behind the actual internal pH changes. Since, as will be shown later, the internal pH is an important condition determining the duration of tension latency, it follows, in agreement with the observations of FIGURE 4, that the measured pH changes would lag behind the latency measurements.

We will now consider the small quickly reversed acidifications that are superimposed on the general pH curve. Dubuisson (1939) has shown that these acidifications, which are attributed to ATP breakdown, occur early in any tetanus, coinciding, in general, with the rise of

tension and then are subsequently reversed. In other words, a decrease of pH results from a slight amount of activity. Since it has been shown for the general effects of the seriate tetani that L_T changes are associated with oppositely directed pH changes, slight activity should be accompanied by an increase in the L_T values. This could not be demonstrated in the tetanus-series type of experiment, since the latency behavior was recorded only for the initial response of each tetanus and not continuously throughout each tetanus as were the pH changes.

The twitch-series experiment, however, permits recording the gradual changes due to continued activity, and reference to FIGURE 3 and PLATE 3 proves, in accordance with the expectation stated above, that after initial slight activity when the pH would be decreased, the L_T values are increased. The experiment of FIGURE 3, although furnishing further confirmation of the correlation between pH and L_T changes, is limited, since its one burst of activity is comparable to only the first tetanus of the experiments summarized in FIGURE 4. The following experiment was therefore performed. A muscle was subjected to a series of bursts of activity, each of which was broken up into two phases: a first phase of six maximal twitches at 3 sec. intervals and an immediately following second phase consisting of a 2 sec. tetanus. Each activity period was followed by a 5 sec. rest. The latency behavior of each of the twitches was recorded, as well as the usual very early response of the tetanus. This procedure, obviously, does not duplicate exactly the technique of Dubuisson, but it does afford a means for following the first few responses in each of a series of bursts of activity as well as the net cumulative changes caused by these successive activity periods.

The results of such an experiment are shown in FIGURE 5. Only the variations in L_1, again used to typify the L_T set, are plotted. (The results for R and L_R are omitted since these are quite like those already presented in the simple tetanus series experiment of FIGURE 2.) It will be noted that L_1 follows the general course previously established, first falling and then rising. But now, in addition, it is noteworthy that, in association with the 6 twitches beginning each activity period, there is a rise in tension latency which is more or less reversed by the remaining activity of that period. These reversible tension rises correspond to Dubuisson's reversible acidifications. The correspondence is by no means exact, but the disparities may be attributed to differences in composition of the associated activity periods in the manner of plotting the activity procedures on the axis of abscissae, and, probably most important, in the relative fidelity with which the registering of the pH

and the latency variations, respectively, reflect the effect of the activity deep within the muscle fiber. Be that as it may, the fact that the slight activity occurring at the very start of each of Dubuisson's tetani and of our twitch-tetanus activity periods results in a reversible acidifi-

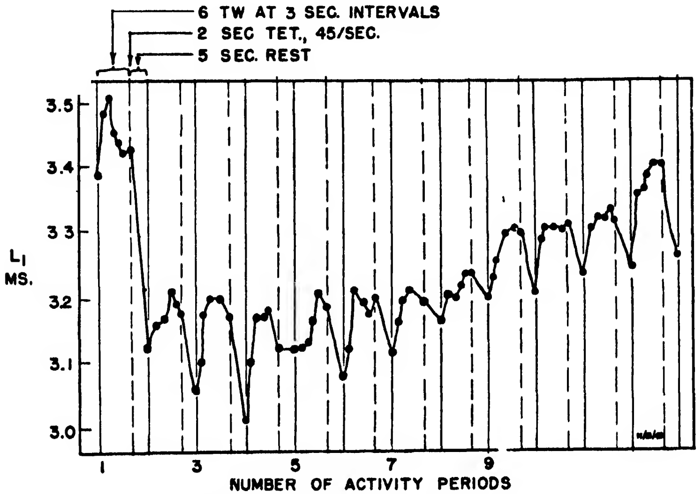


Figure 5 Effect of a series of twitch-tetanus activity combinations on L_I. See text for further details

cation of the muscle in his experiments, and a reversible increase of L_T in ours, is further evidence for the conclusion that the tension latency is smaller, the higher is the pH.

The validity of this conclusion has now been established on the basis of the effects of three different types of activity procedures: (1) the effects of a moderate burst of activity, e.g., a 3 sec. tetanus; (2) the net cumulative effects of a series of 2 sec. tetani; and (3) the gradual changes that occur at the beginning of each of a series of activity units. Brief mention will now be given to the evidence of two other types of experiments. The first duplicates the activity procedure, one twitch every 20 sec. for about two hours, of an experiment by Lipmann and Meyerhof (1930) in which the pH change of a sartorius muscle at a temperature of 20° C. and initially at pH 7.2 was followed manometrically. In our experiment, the L_T's first passed through the usual initial reversible increase during the first few minutes of activity. Thereafter, they gradually decreased for about 40 min. and then increased for the remainder of the run. In the pH experiment, nothing compar-

able to our relatively rapid initial reversible change was recorded since the manometric recording method is too sluggish. But there was recorded a gradual increase in pH for about 50 min. that was then reversed during the remainder of the experiment. It is evident that these changes are in typical correlation with our L_T alterations.

In the second type of experiment, the effect of iodoacetic acid was studied, and it was determined that a 3 sec. tetanus, for example, caused a drop in the L_T's that was not reversed in the subsequent recovery period. In comparable pH experiments, it is found, in consequence of the iodoacetate inhibition of lactic acid formation and the irreversibility of the hydrolysis of creatinephosphate, that irreversible increases in pH occur, hence, again confirming the L_T–pH relation.

Thus, despite the fact that a rather wide variety of activity procedures has been used, the experimentally determined pH and tension latency changes are always consistently related, so that it is possible to conclude, in general, that an increase in pH of a muscle caused by activity is accompanied by a decrease in the tension-latency, or, conversely, a pH decrease, with an L_T increase. (The same general dependence of L_T on the pH has been found in experiments in which the pH was directly modified by immersion of muscles in bicarbonate buffer solutions of different hydrogen ion concentration. For a preliminary report of this work, refer to Sandow, 1943.) Furthermore, this relation seems quite unique, since none of the other variables studied here, L_R, R, or developed tension, exhibit any such regular correlation to reaction changes. The possible significance of this relation will be discussed in the next section.

DISCUSSION

The analysis of the mechanical latent period of contraction given in paper I, the chief results of which are outlined in the introduction to the present paper, proves that at least three more-or-less independent processes occur in this period. The first of these, the latency relaxation induction process, occurs during the initial mechanically quiescent part of the latent period and, in our work, it is measured by the time interval L_R. Nothing is definitely known concerning the detailed nature of this process, although its temporal localization in muscular response suggests that it must involve a reaction, probably the release of Ca which acts as an activator for myosin-adenosinetriphosphatase, that links the action of the stimulus applied to the muscle fibers with the contractile system. In any case, in view of the current results, ac-

tivity, unless carried to states of extreme fatigue, does not affect this reaction.

The other latent period reactions are the LR process, which is the actual mechanism that underlies the observable LR, and the tension-induction process, which is the reaction that directly transforms the contractile machine so that it develops tension. It is our view that both of these processes may be attributed to myosin, and, indeed, the evidence reported in paper I suggests that the tension-induction process is the underlying reaction of the LR; that is, the LR is merely an external mechanical sign of myosin in the course of being transformed from a rested into a contracted state. At any rate, the observable effects of these processes are recorded in the way in which the LR, and at least the onset of tension, develop; and, therefore, they are, in some manner, measured by R_0, R, and the L_T's. Thus, from our results, it is evident that activity has a marked influence on the process that induces tension and on the very beginning of actual tension development, or, in terms of the substance involved, on myosin as it changes into a state of tension.

The above view that the LR process and the tension-induction process are different expressions of the same state is, as yet, an inference (see paper I) which, among other things, obviously requires that the tension induction process follow the LR-induction process and not run temporally parallel to it. It is, therefore, of interest that the present evidence is not in favor of the parallel ordering of these reactions; for the constancy of L_R—i.e., the constancy of the time interval during which no tension change occurs—under conditions involving marked and varied alterations in the speed of tension development, strongly indicates that the contractile myosin has not been affected during the interval L_R. The effects of activity, therefore, suggest, in agreement with the earlier evidence, that the LR-induction process not only leads into the LR, but also into the tension-induction process; and this may be taken as further corroboration of the view that the LR is a mechanical sign of the tension-induction process.

In paper I, a hypothesis was put forward that the tension-induction process is specifically the coupling of myosin and adenosinetriphosphate (ATP) in the form of an enzymatic intermediary complex and that, during this coupling, concomitant with the hydrolysis of ATP to adenosinediphosphate and phosphate, myosin appropriates the energy released by the breakdown of its substrate, and, so energized, contracts. This mechano-chemical mechanism of the nature of the initiation of contraction was based on (1) the assumption that the precontractile

LR is an *in vivo* expression of the important finding of Engelhardt, Ljubimova and Meitina (1941) that artificially spun myosin fibers undergo an increase in extensibility, i.e., they relax, in specific reaction to their catalytic action of splitting adenosinetriphosphate, and (2) the generally accepted view (e.g., Needham, 1942) that ATP is the immediate source of energy for the contractile machine. A prediction of this hypothesis is that the rate of tension development is directly dependent on the rate of hydrolysis of ATP. Although the present research does not involve the measurement of general rates of tension development, the tension latency measurements offer a means of testing this prediction. These give information concerning, at least, the very earliest moments of tension rise, and it seems fair to assume that the more rapidly tension rises, the shorter would be the tension latencies. We would, therefore, expect decreases in the tension latencies under conditions involving an increase in velocity of ATP hydrolysis.

It is, therefore, pertinent to the present work to note that the rate of ATP breakdown is markedly affected by pH. Lehmann (1936) working with muscle mash, Bailey (1942) studying purified myosin-adenosinetriphosphatase, and DuBois and Potter (1943) using ATP-ase extracted from liver, have demonstrated that the rate of this reaction increases with increase in alkalinity up to about pH 9.0. Thus, if the rate of ATP hydrolysis determines the speed of onset of tension, increase of muscle pH should be accompanied by shorter tension latencies. This is precisely what is observed. It is noteworthy, further, that activity does not affect the duration of L_R. Thus, the process that is accelerated by a rise in pH must be localized in the part of the latency response that follows the interval L_R. This fact is also in agreement with our hypothesis, for we have assumed that myosin and ATP do not begin to react until the interval L_R is terminated.

The above discussion envisages a mechanism in which pH affects the rapidity of onset of tension by influencing the rate of decomposition of ATP which is supposed to yield the energy for contraction. It seems that no other muscle chemical reaction could assume this role. This is true, in part, because, in addition to the relation of pH already discussed, no other reaction has the highly important characteristics of being the first chemical transformation of the contractile system to follow stimulation, of having the intimate relation to the contractile machine of being enzymatically catalyzed by the contractile protein, myosin, and of releasing a large supply of energy. Lipmann (1941) has suggested that creatinephosphate might play the part of immediately supplying energy to some biological processes. Such a role for

creatinephosphate is definitely excluded in the direct energizing of muscular contraction, since the rate of breakdown of this substance decreases with increase in pH (Meyerhof and Lohmann, 1928) and, thus, the assumption that creatinephosphate hydrolysis is the immediate source of energy for contraction would involve the inference that the tension latency will decrease with pH increase, which is in contradiction to our experimental results. Furthermore, it would seem that the energy of the splitting of creatinephosphate is not available for directly energizing myosin since, as it is now generally agreed (e.g., see Meyerhof, 1944), this energy is used for the resynthesis of ATP. It is hardly necessary to consider other reactions, such as glycolysis, in relation to our pH-tension latency correlation, since it is clear that, in so far as they release energy, this can be accounted for in other processes than the energization of myosin It, therefore, seems that, among the chemical reactions known to occur in muscle, the hydrolysis of ATP is unique in possessing many special properties that are in accord with the view that it is directly involved in the process that energizes and induces tension development of myosin in stimulated muscle.

The effect of pH on the early rate of tension development might still be mediated in part by an influence on the physical properties of myosin that determine its contractility. In this regard, it is significant that Dainty *et al* (1944) have shown that pH changes over the range from 6.0 to 8.5 do not affect the flow-birefringence or the relative viscosity of myosin sols. Our pH differences due to activity fall well within this range, and if it is possible to apply results concerning myosin sols to the myosin fibers of live muscle, then the flow-birefringence and relative viscosity of myosin, whatever their relation to contractility, would, in virtue of their constancy, not be responsible for the observed latency changes.

Von Muralt (1934), however, by measuring changes in transmission of light by muscle, concludes that activity does cause some physicochemical changes in the state of the muscle protein, presumably myosin. These changes are related to the breakdown and resynthesis of creatinephosphate and not to the pH alterations. This is important, for although some aspects of the pH alterations are attributed to creatinephosphate reactions, not all of them can be so explained. The acidification that accompanies slight activity is due to the splitting of ATP, and the decrease in pH that occurs relatively late in a prolonged activity series is due principally to the accumulation of lactic acid. Hence, the changing rate of rise of tension cannot be correlated to creatinephosphate reactions alone, and this definitely indicates that the changes

in muscle proteins observed by von Muralt do not involve changes in the contractility of the myosin which could be correlated to our changes in tension rise. Therefore, as far as the above evidence is concerned, pH alterations do not affect the development of tension by means of some direct influence on myosin.

It is now evident that the observed relation between pH and the tension latencies is consistent with the assumption that the energization of myosin for contraction is dependent on the hydrolysis of ATP during the tension induction process, and furthermore, that this relation, in so far as the various researches discussed here shed light on the matter, is not explained by other means. To this extent, the evidence, therefore, indicates that the mechanism of interaction of myosin and ATP we have assumed to account for the energization of contraction is correct.

It is important to note, however, that the pH-L_T correlation has significance independently of the feature of this mechanism embodied in the view that the LR is a mechanical sign of an enzymatic intermediary combination between myosin and ATP. Keeping in mind the unique set of properties of ATP in the energetics and chemistry of muscle metabolism already enumerated, it is clear that the fact that a change in muscle pH causes the rates of tension onset and ATP hydrolysis to vary in the same sense, strongly suggests that ATP not only energizes myosin for contraction, but, in the sense of Dainty *et al.* (1944), is also the agent that, upon receipt of a suitable stimulus, directly activates the myosin to contract. Additional support for this view is found in the recent studies of Buchthal and his coworkers (1944a, 1944b) These investigators have shown that the application of small amounts of ATP, either by means of close arterial injection to mammalian skeletal muscle or directly to mammalian smooth muscles and to single amphibian skeletal muscle fibers, always results in the release of contractions accompanied by action potentials in all of these muscles. These responses are obtained even in curarized muscle fibers and in atropinized smooth muscle, thus proving that the action of the ATP is directly upon the muscle fibers involved, and not upon some other feature of the relevant neuromotor complexes. It does not seem to have been demonstrated that the contact of the ATP with the muscle fibers has its primary effect on the contractile mechanism and not on the excitatory structures. The results, especially in view of the responses of the curarized and atropinized preparations, nevertheless strongly indicate that the applied ATP does act directly on the contractile system by energizing and activating the myosin to contract.

In view of the evidence of the present research and the concordant results of Buchthal and his coworkers that ATP acts as the activating agent for contraction, it is of interest to discuss the studies of Ritchie (1933). This investigator raised the question that energy-coupling between the immediate source and the contractile machine might conceivably occur in either of two temporal possibilities: (1) during some phase of the general recovery process, and this he designated the "physical theory," or (2) during the latent and contraction periods, the "chemical theory."[1] Some discussion of this problem has recently appeared in the work of Kalckar (1938), Bailey (1942), and Dainty *et al.* (1944). Clearly, our conception of energy-coupling in muscular activity and Buchthal's, also, are in accord with Ritchie's "chemical theory." Ritchie's experiments, however, led him to conclude that the "physical theory" was true. His conclusion was based on a series of experiments on frog heart muscle involving the effect of pH on the latent and refractory periods. His latency results are in agreement with ours. But, in interpreting his results, Ritchie accepted the view prevalent in 1933 that creatinephosphate was the immediate source of energy for the contractile substance. Since, as already pointed out, the rate of hydrolysis of this substance decreases with increase in pH, it is clear that he was forced to the conclusion that this reaction could not energize the muscle in a latency process which was quickened by an increase of pH, and that, therefore, the energization must occur in a recovery reaction. A corresponding analysis of the refractory period results led to the same final conclusion.

We have no disagreement with Ritchie's experimental results nor with the general form he adopts for their interpretation. It is evidently necessary, however, to reinterpret his experiments, since, today, ATP and not creatinephosphate is known to be the substance that directly supplies energy to myosin. Recalling our discussion and the fact that Ritchie's results, like ours, prove that the mechanical latency is shorter the higher the pH, it is then clear that his findings support the chemical theory of contraction. Thus, his experiments and ours lead to the common conclusion that the transfer of energy to the contractile substance occurs immediately after the application of the stimulus, and that this energizing process simultaneously activates the myosin to contract.

[1] Some criticism (see Dainty *et al.*) has been made of Ritchie's terminology, on the ground that the terms, "chemical" and "physical," are ambiguous. We agree with this criticism. We wish further, to point out that the essential issue is whether the chemical that immediately supplies energy to myosin, directly activates, as well as energizes it for contraction (chemical theory), or merely energizes it in some recovery process (physical theory). It is, therefore, suggested that the terms, "activation coupling" and "recovery coupling," be used instead of, respectively, "chemical theory" and "physical theory."

This conception of the mechanism by which contraction is initiated is, in certain respects, comparable to the alpha-process Brown (1941) has postulated for the activation of myosin for tension development. Indeed, it is possible, that what we have called the tension-induction process is identical with the alpha-process, and thus it follows that the latency relaxation would be the part of the externally recorded mechanical response of muscle that directly represents the alpha-process. Some attention has been given to this point, theoretically, in a preliminary notice (Sandow, 1945) of a mathematical formulation of our hypothesis of the energizing mechanism. To be especially noted here, however, is the fact that Brown's and our views are in agreement, at least in placing the events that are involved in the energization and activation of myosin for its mechanical activity in the short period of time immediately following the application of the stimulus to the muscle.*

It must be emphasized, however, that our own evidence for the particular mechanism we have proposed for the tension-induction process is essentially based only on certain features of the latency response Thus, in our work, no attempt has been made to interpret the variations in R due to activity. There are indications that R varies directly with the concentration of creatinephosphate and this suggests that the formation of the LR may be dependent on this substance, probably indirectly, in relation to the function of creatinephosphate as an energy reservoir for ATP (see Lipmann, 1941). At any rate, the fact that, in general, activity causes R to vary in a quite different manner than the L_T's, leads us to conclude that the depth of the LR is determined by other mechanisms than those proposed to account for the speed of the latent period processes that cause tension to develop. It should also be noted that no attempt is made in this paper to relate the latency behavior with the peak tension output, records of which are included in our results, nor with other features of the entire contraction, such as the general rates of tension change during the contraction and relaxation periods. Studies of this type are clearly important in checking the validity of conceptions of tension development based principally on latency phenomena. It is our intention to discuss such problems in later publications concerned with research now in progress (for a preliminary report, see Sandow, 1944a) on the various features of the entire twitch.

* The above paragraph was inserted subsequent to the award of the A Cressy Morrison Prize.

SUMMARY

1. Studies have been made of the effect of activity, such as a tetanus, a series of tetani, and a series of twitches, on the mechanical events of the latent period of frog skeletal muscle. These events, specially characterized by the minute pre-contractile latency relaxation (LR), are recorded by a piezo-electric, cathode-ray oscillographic technique which, with great reliability, permits determining the activity induced latency time differences of the order of several 0.1 ms. and length differences of about 0.05 μ.

2. Activity in any amount, even a single twitch, causes a reduction in R (the final depth of the LR), and the greater the activity, the greater is the decrease in R.

3. L_R, the time interval between the instant of stimulation and the beginning of the LR, is not affected by activity up to quite advanced fatigue. Beyond this, activity causes L_R to increase.

4. The three members of the set of latencies for tension development, collectively symbolized by L_T, change in parallel, in consequence of any activity sequence. The L_T's, first, increase with slight activity (5 or 6 successive twitches); then decrease with continued activity (e.g., up to 4 or 5 successive 2 sec. tetani); and then increase again with further continuation of the activity.

5. The listed changes in the latent period variables due to activity, are in general reversed by rest.

6 Special control tests prove that the changes in the latency behavior, which is an expression of the response of the muscle's contractile system, are due to direct effects of the activity on the contractile system, and not to effects on the excitatory system, which are merely reflected in contractile behavior.

7. The L_T changes are correlated in some detail with activity induced pH changes, studied in other investigations, so as to prove that the higher the pH, the shorter the L_T's.

8. This correlation and several other implications of the data are shown to be confirmatory evidence for a previously advanced hypothesis that the LR is a mechanical sign of a tension-induction process involving mechano-chemical coupling of myosin and adenosinetriphosphate in the form of an enzyme-substrate combination, during the existence of which the myosin is energized and activated for contraction.

LITERATURE CITED

Bailey, K.
1942. Myosin and adenosinetriphosphatase. Bioch. J. **36**: 121-139.

Brown, D. E. S.
1941. The regulation of energy exchange in contracting muscle. Biol. Symposia **3**: 161-190.

Buchthal, F., A. Deutsch, & G. G. Knappeis
1944a. Adenosinetriphosphate initiating contraction and changing bi-refringence in isolated cross-striated muscle fibers. Nature **153**: 174.

Buchthal, F., & G. Kahlson
1944b. Application of adenosinetriphosphate and related compounds to mammalian striated and smooth muscle. Nature **154**: 178-179.

Dainty, M., A. Kleinzeller, A. S. C. Lawrence, M. Miall, J. Needham, D. M. Needham, & Shih-Chang Shen
1944. Studies on the anomalous viscosity and flow-birefringence of protein solutions. III changes in these properties of myosin solutions in relation to adenosinetriphosphate and muscular contraction. J. Gen. Physiol. **27**: 355-399.

Dubuisson, M.
1937. Untersuchungen über die Reaktionsänderung des Muskels im Verlauf der Tatigkeit. Pflüg. Arch. **239**: 314-326.
1939. Studies on the chemical processes which occur in muscle before, during and after contraction. J. Physiol. **94**:461-482.

DuBois, K. P., & V. R. Potter
1943. The assay of animal tissues for respiratory enzymes. III. Adenosinetriphosphatase. J. Biol. Chem. **150**: 185-195.

Engelhardt, W. A., M. N. Ljubimova, & R. A. Meitina.
1941. Chemistry and mechanics of muscle studied on myosin. Compt. Rend. Acad. Sci. U.S.S.R. **30**: 644-646 (as abstracted in Chem. Abst. **37**(2): 391-392. 1943).

Hill, D. K.
1940. Hydrogen-ion concentration changes in frog's muscle following activity. J. Physiol. **98**: 467-479.

Kalckar, H. M.
1941. The nature of energetic coupling in biological systems. Chem. Rev. **28**: 71-178.

Lehmann, H.
1936. Ueber die Umersterung des Adenylsäuresystems mit Phosphagen. Bioch. Zeitschr. **286**: 336-343.

Lipmann, F.
1941. Metabolic generation and utilization of phosphate bond energy. Advances in Enzymology **1**: 99-162.

Lipmann, F., & O. Meyerhof
1930. Ueber die reaktionsänderung des tätigen Muskels. Bioch. Zeitschr. **227**: 84-109.

Meyerhof, O.
1944. Energy relationships in glycolysis and phosphorylation. Ann. N. Y. Acad. Sci. **45**: 377-393.

Meyerhof, O., & K. Lohmann
1928. Ueber die natürlichen Guanidinophosphosäuren (Phosphogene) in der quergestreiften Muskulatur. I. Das physiologische Verhalten der Phosphogene. Bioch. Zeit. **196**: 22-48.

Meyerhof, O., W. Mohle, & W. Schulz
1932. Ueber die Reaktionsänderung des Muskels im Zusammenhang mit Spannungsentwicklung und chemischen Umsatz. Bioch. Zeitschr. **246**: 285-318.

Millikan, G. A.
1942. The chemistry of muscle. Ann. Rev. Bioch. **11**: 497-510.

Needham, D. M.
1942. The adenosinetriphosphatase activity of myosin preparations. Biochem. J. **36**: 113-120.

Rauh, F.
1922. Die Latenzzeit des Muskelelementes. Zeit. f. Biol. **76**: 25-48.

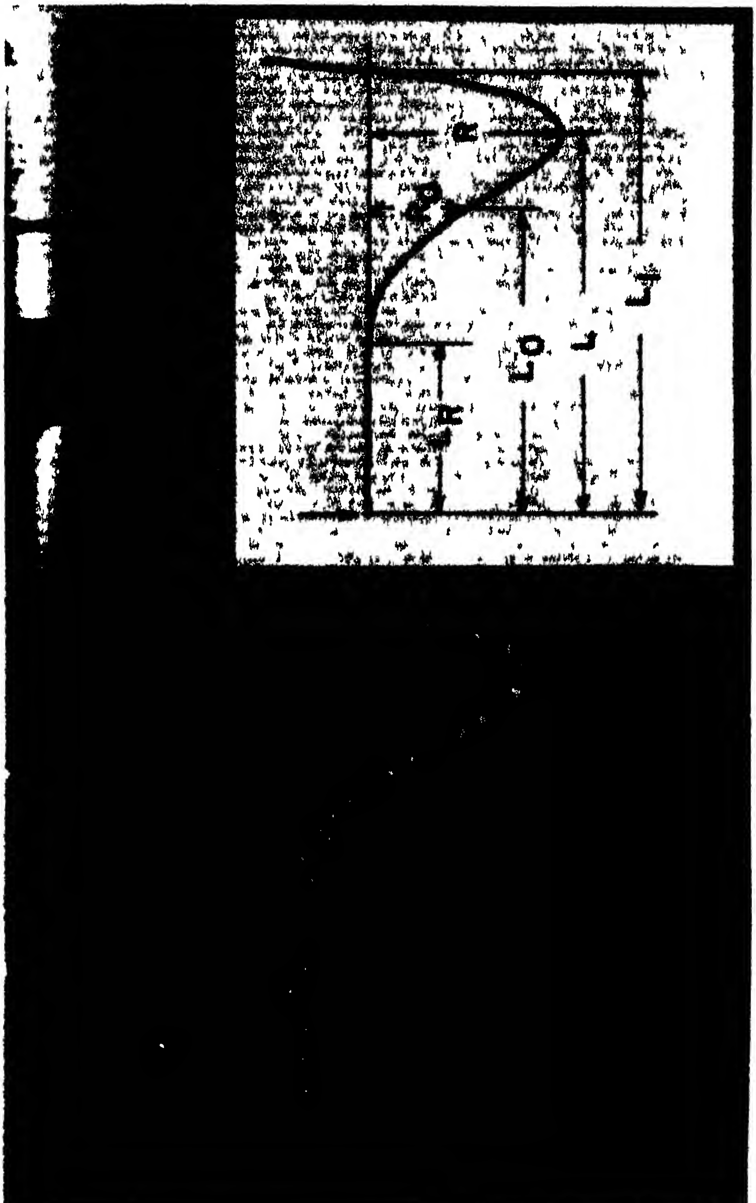
Ritchie, A. D.
1933. Theories of muscular contraction. J. Physiol. **78**: 322-334.

Sandow, A.
1942a. Mechanical changes of muscle during the latent period of isometric contractions. Federation Proc. **1** (1): 77.
1942b. The effect of tetanus on muscular latency relaxation in normal and iodoacetate-poisoned muscles. Anat. Rec. **84**(4): 21-22.
1943. Study of the effect of pH, tissue poisons, and anistonicity on the mechanical events of the contraction and relaxation periods of skeletal muscular contraction. Year Book of the Amer. Philos. Soc.: 195-198.
1944a. (with the technical assistance of A. G. Karczmar). The effect of activity on the twitch of frog skeletal muscle. Anat. Rec. **89**: 17-18.
1944b. Studies on the latent period of muscular contraction. Method. General properties of latency relaxation. J. Cell and Comp. Physiol. **24**: 221-256.
1945. The mechanism of energizing muscular contraction. Trans. N. Y. Acad. Sci. (II) **7** (6): 139-151.

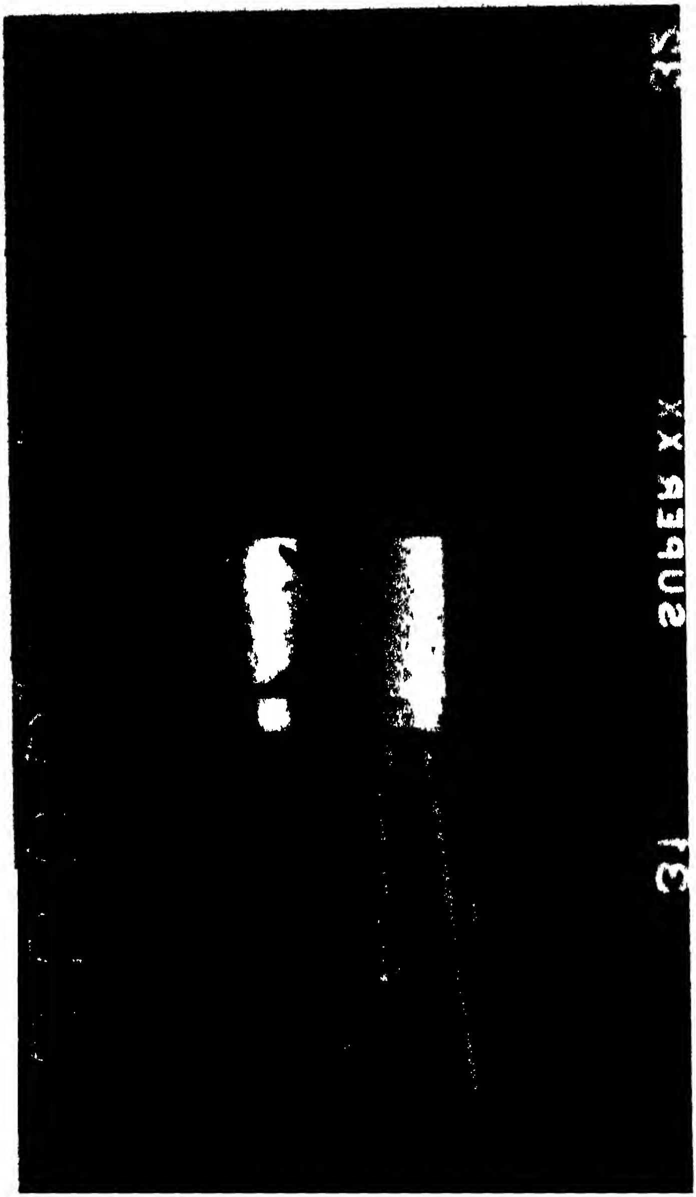
von Muralt, A.
1934. Lichtdurchlässigkeit und Tätigkeitsstoffwechsel des Muskels. II. Mitteilung. Pflüg. Arch **234**: 653-664.

PLATE 1

Typical record of a latent period response of a sartorius muscle. The sinusoidal modulation of the record is a 10,000 cycles/sec timing wave The inset gives the symbols of the measured variables. The L's are latency time intervals, all measured from the instant of stimulation (indicated by the arrow), as follows: L_R, to the beginning of the latency relaxation (LR); L_e, to the point of inflection of the LR curve, L, to the deepest point of the LR; and L_1, to the instant at which the latency mechanogram just crosses the original resting-tension base-line The R's represent decreases in tension, or equivalent increases in length, of the muscle at the indicated LR points The light band, at the upper right, measures the initial and peak developed tension of the contraction $\times \frac{1}{2}$.



SANDOW ACTIVITY EFFECTS ON LATENCY MECHANISMS



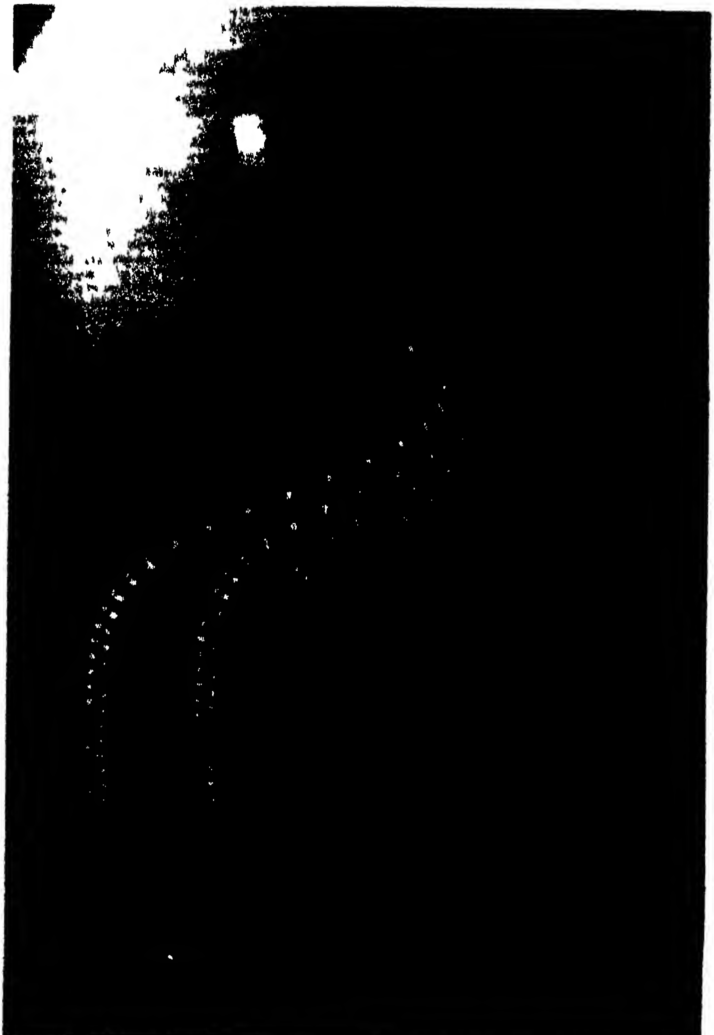
SANDOW ACTIVITY EFFECTS ON LATENCY MECHANISMS

PLATE 2

Immediate effect of the 3 sec tetanus of FIGURE 1 (page 159) on the lateney response Left record shows the latency response of the first twitch of the tetanus The right record exhibits the latency behavior 15 sec after the tetanus was applied Note decreases in L and R $\times {}^2/_3$.

PLATE 3

Record showing the first 5 latency responses of the twitch-series activity experiment plotted in FIGURE 5. The first response is at the top of the set, the others follow in order downward. 10,000 cycles/sec timing wave. Note in the successive responses, the gradual increase of L and the decrease in R. $\times 2/3$



SANDOW ACTIVITY EFFECTS ON LATENCY MECHANISMS

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES

VOLUME XLVI, ART. 4. PAGES 185–208

AUGUST 18, 1945

RESPIRATION AND GERMINATION STUDIES OF SEEDS IN MOIST STORAGE*

BY

LELA V. BARTON†

CONTENTS

* Awarded an Honorable Mention in the A. Cressy Morrison Prize Competition in 1944. Publication made possible through a grant from the income of the Nathaniel Lord Britton Fund.

† Boyce Thompson Institute for Plant Research, Inc., Yonkers, N. Y.

INTRODUCTION

Numerous investigators have been concerned with the viability of seeds buried in the soil for long periods of time. Weed seeds especially have been the subject of many studies. No effort will be made here to give a complete review of the published literature on this subject. One experiment which was started by Beal will serve to indicate the remarkably long life span of certain weed seeds buried in the soil. Twenty sets of seeds were buried in 1879 in uncorked bottles in sandy soil 20 inches deep with the mouths tilted downward to avoid filling with water. *Amaranthus retroflexus* and *Rumex crispus* were among the seeds which remained viable for 40 years and some *Rumex crispus* seeds still germinated after 60 years' burial.[6] Since the seeds must have absorbed water soon after being placed in the soil and must have remained imbibed with water throughout the storage period, this length of life is noteworthy. Furthermore, these seeds are not usually considered dormant. That they do not develop a deep dormancy, when buried in the soil, is demonstrated by the fact that when the soil in an ordinary garden plot, for example, is disturbed by cultivation, many weed seedlings appear even if care has been taken to remove all such plants for several preceding years.

When seeds take up water preparatory to germination, their metabolic activities become greatly accelerated. It is evident that imbibed seeds could not remain viable very long in the soil unless definite curtailment of these activities take place. The present respiration studies were undertaken in an attempt to get a definite measure of one of the life activities of seeds which were imbibed with water but which were kept from germinating by placing at unfavorable temperatures. Studies were also made of the germination behavior of such seeds.

RESPIRATION

Materials and Methods

Seeds of *Amaranthus retroflexus* L. and *Rumex obtusifolius* L. were selected as experimental material, because some work had been done on these genera previously and because they were readily available for collection. Limited tests were also made using *Impatiens balsamina* L. seeds produced on the Institute grounds.

After collection the seeds were cleaned of all extraneous material and then counted in lots of 2500 each for *Amaranthus*, 1000 each for *Rumex*,

and 100 each for *Impatiens*. A vacuum counter was used for *Amaranthus* seeds. Counted instead of weighed lots were chosen since it has been demonstrated that the number of individual embryos is important in determining respiration. Bailey and Gurjar[1] pointed out that shriveled wheat respired at a rate two or three times greater than did plump grains. This they attributed to the higher ratio of embryo weight to endosperm weight in shriveled grains as compared with that of the plump wheat. It has been demonstrated, however, that some endosperm tissues are capable of respiration.[15]

The counted lots of dry seeds were remarkably uniform in their weights. For example, weights of five lots of 2500 seeds each of *Amaranthus* were 0.94, 0.96, 0.95, 0.96, and 0.96 gram; of five lots of 1000 seeds each of *Rumex* 1.21, 1.22, 1.21, 1.23, and 1.25 grams; and of five lots of 100 seeds each of *Impatiens* 0.49, 0.51, 0.53, 0.45, and 0.57 gram. Variations in the weights of *Impatiens* were due to different sizes of seeds in the lot used, in spite of the fact that the largest and smallest seeds were screened out.

Since respiration measurements were to be made, care had to be taken to eliminate environmental factors which would alter gaseous exchange The seeds were kept in constant temperature rooms large enough so that all respiration measurements could be made there. As will be shown later, imbibed seeds are very sensitive to temperature changes Furthermore, mere mechanical disturbance of such seeds may bring about germination, in which case respiration would be greatly increased. This posed the problem of getting the seeds into a respiration chamber without disturbance. The appearance of mold on the seeds would also alter respiration, as might any sterilizing agent which could be used.

With all of these factors taken into consideration and after a considerable number of trials, the following method was adopted. The seeds were cleaned thoroughly by compressed air, but not washed. They were then placed on strips of moist glass wool. Glass wool was used because it tended to decrease the possibility of infection. It held moisture well, and was easy to handle. These strips of glass wool with the seeds were placed in sterile Petri dishes where they were kept until the respiration measurement was made. Constant temperatures used to hold moist seeds without germination were 20° C. for seeds of *Amaranthus* and *Impatiens*, and 30° C. for seeds of *Rumex*. They were examined at intervals and the moisture adjusted, using boiled tap water, and any seedlings removed. Also, careful check was kept on possible mold formation and any seed showing the slightest sign of infection was removed and discarded. A record of germinations and moldy discards

was kept for each culture so that the exact number of seeds used for each respiration test was known.

For measurement of respiration, the glass wool strip containing the seeds was removed from the Petri dish and placed on a strip of paraffined cheesecloth of similar size. This was then inserted in the respiration tube as shown in PLATE 1. This tube was fitted at one end with a ground glass stopper the size of the internal bore of the tube and, at the other end, with a capillary stopcock and tubing to facilitate the removal of the gas sample. Usually, three to five tubes were set up for each test, since a culture could not be used for respiration measurements if seedlings or mold had appeared by the end of the respiration period. After several preliminary tests, a 48-hour respiration period was shown to be satisfactory.

The volume of gas in the respiration chamber varied according to the tube used and the size of the seeds, but was usually between 65 and 75 ml. Since this was 25 to 40 or more times the volume of the seeds being tested, respiration for 48 hours was permitted without unduly reducing the oxygen content or increasing the carbon dioxide content. At the same time, sufficient change occurred in the proportion of the two gases to give satisfactory measurements by gas analysis.

It is recognized that the moisture in the respiration tube might have absorbed some of the respired gases Some measurements of oxygen and carbon dioxide in control tubes, with and without moist strips of glass wool, indicated a decrease in the former of 0 06 per cent CO_2 and 0.2 per cent O_2. This means then that all of the respiration readings for seeds were slightly lower than actual, but the constancy of the error permitted comparisons to be made.

The necessity of having a separate air control for each test was demonstrated early in the experimental work, for the air in the constant temperature rooms varied in oxygen and carbon dioxide content from time to time. Consequently, all of the results have been calculated using individual and separate controls for each test. The control tubes contained no moist glass wool.

At the end of the 48-hour period, the gas in the respiration tube was displaced by using a 22 per cent aqueous solution of sodium chloride recommended by Dennis and Nichols[7] as having minimum solubility for both oxygen and carbon dioxide. Analysis of the gas for both oxygen and carbon dioxide was made in an apparatus which is shown in PLATE 2. The apparatus is essentially that described by Haldane and Graham,[8] with certain adaptations and modifications designed by Dr. Norwood C. Thornton of this Institute to fit the special requirements

of these experiments. There was a permanent mounting of the gas analysis and gas sampling and measuring burettes by a T connection, in order to take in samples and expel air from the gas analysis burette without loss of time or errors in handling of the original gas sample. The small absorption pipettes were filled with glass tubes to enlarge the absorption area of the liquid. The pipettes were connected to large reservoirs of the absorbing liquids in order that several consecutive determinations could be made without replenishing the absorbents. Fifteen per cent sodium hydroxide solution was used to remove carbon dioxide from gas samples. Oxygen was absorbed in an alkaline pyrogallate solution made by dissolving 26.25 grams pyrogallic acid in 175 ml. of 51 per cent potassium hydroxide. Solutions were changed when oxygen or carbon dioxide absorption became sluggish. The solutions lasted for at least 20 analyses. The volume of alkaline pyrogallate solution used in the apparatus was 175 ml. and that of sodium hydroxide between 60 and 70 ml.

The sample to be analyzed was displaced from the respiration tube A, PLATE 2, into measuring tube B by manipulation of reservoir L and by displacement of the gas from A with 22 per cent solution of sodium chloride following the removal of all air from the connecting tubes and the removal of the ground glass stopper of A under the surface of the sodium chloride solution. The sample was collected and measured in a gas burette B of 100 ml. capacity, with a graduation of 0.2 ml. This displacement by salt solution over to a tube containing mercury was effected with little or no carry over of water. If some water did go over, it was separated by settling on the mercury, so that an air sample free of water droplets could be drawn into the gas burette C. For gas analysis, 16 ml. were taken into this burette, which is graduated to 0.01 ml. and can be read accurately to 0.005 ml. The graduated part of this tube extends from 11.85 ml. to 16 ml. The remainder of the 16 ml. volume is supplied in the bulb at the top of the tube. Surrounding this gas burette is a water jacket and, attached to it by means of thick-walled rubber tubing, is a leveling vessel of mercury D, which is supported and manipulated by a pinion E. The gas sample was then ready for removal of carbon dioxide, which was accomplished by forcing the gas through the sodium hydroxide solution in absorption pipette F, which was filled with glass tubes to increase the absorbing surface. By raising and lowering the mercury through manipulation of vessel D, the gas was passed over the sodium hydroxide solution seven times and then a reading was taken on gas burette C. The difference between this reading and 16 ml. indicated the amount of carbon dioxide

in the sample. After the reading was taken, the gas was passed over the hydroxide solution four more times and a second reading taken. If the two readings were identical, it was assumed that all of the carbon dioxide had been removed.

Absorption pipette F was then closed off and the same gas sample sent over the alkaline pyrogallate solution in absorption pipette G, also provided with glass tubes, and attached to reservoir H. The gas was sent over the pyrogallate solution ten times and then once again over the hydroxide solution to wash out the oxygen in the connecting tube. Then, sending it five additional times through the pyrogallate solution, followed by another washing out of the connecting tube, usually removed all the oxygen of the sample. After adjusting the pressure in the burette by using the hydroxide pipette, which had a scale attached on the capillary portion, and which had the other arm attached to the compensating tube K of the same design as the gas burette and also inserted in the water jacket, a reading was taken on gas burette C. A reading of the amount of oxygen in the sample could then be had by subtracting this value from the carbon dioxide reading already obtained. These values could be converted to percentages by multiplying the number of milliliters of carbon dioxide or oxygen by 6.25, as the total amount of the sample was 16 ml. Since the limit of reading on the gas burette was 0.005, the limit of accuracy in per cent was 0.03125.

At the beginning of each test, any pressure due to different levels of the pyrogallate solution and the water seal was released by opening a stopcock, between the two liquids, to the air. Also, the pressure on the sodium hydroxide solution and in the connected capillary compensator tube K was made the same as that of the atmosphere by opening the stopcock (above the letter K, PLATE 2) to the air. Before a final reading was taken, after any absorption, the gas was always passed through the alkali tube in which compensation had been made for the atmospheric pressure, and the level of the liquid in the tubing below the alkali stopcock had been brought to a predetermined level, set with the stopcock of the compensator tube open to the air before the gas analysis was started. Since the analyses were always made within a closed system, and pressure adjustments made before each test, barometric pressure effects may be disregarded in the results. Temperature effects were eliminated by the use of constant-temperature rooms.

The carbon dioxide and oxygen were always removed from the air within the gas-analysis apparatus before a new sample was taken, so that carbon dioxide- and oxygen-free air occupied the connecting tubes between G, F, and C. The connecting tubes between C, B, and A, as

well as the small connecting tube on the respiration chamber itself (shown in PLATE 1), were filled with mercury before the gas was drawn from A.

Tests to determine the reproducibility of the respirometer readings were made. From a sample of air taken directly into the measuring burette, five tests of 16 ml. each were made. Duplicate readings from each test gave identical results for four of the five tests. The fifth test showed 0.02 per cent more carbon dioxide and 0.02 per cent less oxygen. This was within the limits of accuracy of the apparatus, which was 0.03125 per cent. Three tests on a sample of air taken from a respiration tube into the apparatus, with duplicate readings on each test, gave identical results for carbon dioxide and a difference of 0.02 per cent, in one case, for oxygen. A repetition of these tests on an air sample from another respiration tube gave identical amounts of carbon dioxide and oxygen in all cases. Also different operators obtained the same measurements from two samples of the same gas mixture.

Results and Discussion

Amaranthus retroflexus. Respiration measurements were begun, in the fall of 1941, on freshly-harvested seeds, but no reliable measurements were obtained until the respirometer modification was complete, so early measurements were meaningless. Some of the determinations made on these seeds at later periods will be given below.

Two seed collections were made in the fall of 1942. Collection A was obtained in Yonkers on September 28. Forty counted lots, of 2500 seeds each, were placed on moist glass wool in a room at 20° C. on October 6, 1942. Collection B was made at Emmaus, Pa., and shipped directly to Yonkers, arriving in this laboratory on October 1, 1942. Lots, as for Collection A, were placed on moist glass wool in a 20° C. room on October 13, 1942. The carbon dioxide output and the oxygen intake of these seeds after various lengths of time in moist storage are shown in TABLE 1. The gaseous exchange is expressed as milliliters of carbon dioxide evolved and oxygen absorbed by 2000 seeds in a 48-hour period.

It should be kept in mind that 20° C. is an inhibitive germination temperature for dry seeds only when they are freshly harvested. As they after-ripen in dry storage they become less exacting in their temperature requirements for germination. Also, crops and collections vary in the degree of dormancy exhibited by freshly-harvested seeds. Seldom does one find a collection in which all of the seeds are dormant. It was therefore, difficult to get respiration measurements of dormant

TABLE 1

RESPIRATION OF SEEDS OF *AMARANTHUS RETROFLEXUS* HELD ON MOIST GLASS WOOL AT 20° C. FOR DIFFERENT LENGTHS OF TIME AFTER HARVEST

(Carbon Dioxide Given Off and Oxygen Taken up Expressed as Ml. per 2000 Seeds for a Period of 48 Hours)

Collection A				Collection B			
Days in moist storage	CO_2	O_2	CO_2/O_2	Days in moist storage	CO_2	O_2	CO_2/O_2
0	1.52	2.20	0.69	0	1.42	1.94	0.73
1*	1 74*	2.80*	0.62	1	1.36	1.92	0.71
2	1.36	2.20	0.62	2	1.18	1.76	0.67
8	0.90	1.48	0.61	8	0.66	1.16	0.57
16	0.62	0.98	0.63	16	0.04	1.18	0.03
35	0.48	0.94	0.51	32	0.62	1.00	0.62
67	0.34	0.76	0.45	63	0.38	0.84	0.45
134	0.32	0.66	0.48	134	0.26	0.60	0.43
218	0 40	0.80	0.50	211	0.40	0.62	0.65
385	0.18	0.68	0.26	375	0.24	0.50	0.48
564	0.18	0.42	0.43	551	0.12	0.98	0.12

* 3 small seedlings + 2 seeds split at end of respiratory period

seeds for the one- and two-day moist storage periods. In spite of the fact that ten replicate tubes were set up for each seed collection for each of these periods, a reading for dormant seeds of collection A, after one day of storage, was not obtained. All of the tubes contained some seedlings. The best one had three small seedlings and two additional seeds split at the end of the respiration period. These data have high values as can be seen in the table. All of the other values were fairly consistent with the exception of the high oxygen uptake value for collection B, after 551 days of storage. Experimental procedure failed to reveal the cause of this.

It is evident that the respiration of seeds which imbibe water and still remain dormant was reduced soon after the water was absorbed. Between two and eight days, a rather sharp drop in respiratory activity was noted. This decrease continued with increased length of the storage period and probably was still to be seen after 551 and 564 days of moist storage. That the amounts of carbon dioxide given off and oxygen taken up will continue to diminish in these seeds is indicated by the data obtained from seeds of the 1941 crop. Respiratory activity of these seeds was measured after 375, 728, and 901 days, when the carbon dioxide given off per 2000 seeds, in a 48-hour period, was found to be 0.08, 0.02, and 0.04 ml. for the three periods respectively. The oxygen consumption for the same periods was 0.66, 0.36, and 0.28 ml. These seeds of the 1941 crop, stored 901 days, were still viable, as will be shown by germination tests below.

The respiratory quotient tended to decrease with increased time in moist storage (TABLE 1). The importance of the respiratory quotient as an index of what actually occurs in gaseous exchange of materials has been questioned by some workers. Many others, however, believe that certain conclusions can be drawn from the relative amounts of carbon dioxide given off and the oxygen taken up. Certainly, it can be said that the conditions under which the seeds were kept in this experiment affected not only the absolute, but also the relative rates of the two processes concerned (TABLE 1).

Sherman[13] conducted a study on the respiration of dormant seeds which included those of *Amaranthus retroflexus*. Measurements were made at intervals from 3 to 176 days after harvest. These seeds were held dry, however, until the time of testing, so that they did not remain dormant. She obtained variations in respiration in the different lots tested, but the respiratory quotient remained the same throughout. This value was somewhat higher (0.8) than that obtained with freshly-harvested seeds in the present study, but was still less than unity.

One of the chief determining factors in the value of the respiratory quotient is the chemical composition of the respiring materials. A chemical analysis of *Amaranthus retroflexus* seeds made by Woo[17] showed, that of the total constituents, 47.03 per cent are carbohydrates and 7.86 per cent lipins. From these data, one would expect a respiratory quotient approximating unity. On the contrary, the actual values found for freshly-harvested seeds was approximately 0.7 which is the respiratory quotient for the absolute oxidation of fats. At any one time, the respiratory quotient is related to the nature of the substance actually used at that time. Since the seeds contain very little fat, it is probably concentrated in the embryo. It is possible, then, that this oil supply is drawn upon immediately when the seed absorbs water and much oxygen used in its conversion to sugar. This would account for the initial low value of the respiratory quotient. Stiles and Leach,[14] reporting the respiration of germinating *Fagopyrum esculentum* seeds, said that the small reserve of fat in this seed was consumed at a very early stage in seedling development, in spite of the fact that the principal food reserve in the seed was starch.

Some tests were made on seeds of collection A, 1942 crop, in July 1944, to determine the respiration upon transfer to a 30° C. room. Respiration measurements were made after 1, 3, 6, 24, and 48 hours in a respiration tube. The seeds were placed in the respiration tube directly upon removal from 20° C. to the 30° C. room. The tubes and the paraffined cheesecloth had been kept at 30° C., but the seeds were

left on the original moist glass wool which had been at 20° C. Respiratory activity could be measured, after one hour, at 30° C. The actual amounts of carbon dioxide given off and oxygen taken up were small, but the respiratory quotient was back up to 0.73 at that time. The amounts of gas respired increased with longer periods, but the quotient remained the same even after 24 hours in the respiration tube, when carbon dioxide and oxygen contents of the gas had reached 5 and 7 per cent, respectively, and when many seedlings had appeared in the tube. These results agree with those of Stiles and Leach[14] that the fat reserve is still being used at an early stage of seedling development. An extension of time in the respiration tube at 30° C. to 48 hours, the period used for all of the tests at 20° C., yielded 24 per cent carbon dioxide, and indicated the absorption of 20 per cent oxygen. The numerous seedlings were quite large at this stage and the considerable quantity of carbon dioxide, coupled with the exhaustion of the oxygen supply, doubtless affected the respiratory rate, so that these data can not be used for comparison. The increase after one hour at 30° C. in the respiratory quotient from the low value of the moist ungerminated seeds to the original value at the time of storage is not surprising in view of the sensitivity of these seeds to temperature changes reported below.

Brown[3] measured the respiration of dormant red oak acorns and found a lowered respiratory quotient with increased time in storage for three weeks at 2.5° C. This was due to an increased oxygen consumption, while carbon dioxide production remained the same. In the present tests, both carbon dioxide produced and oxygen absorbed were decreased with storage, but the proportions did not remain constant. This continued reduction in the value of the respiratory quotient with increased length of time that the seeds were held ungerminated at 20° C. might indicate the use of oxygen for purposes other than respiration. Such a storage of surplus oxygen is known to lead to the accumulation of acids and sugars. Also, it should be noted that a retention of carbon dioxide in the tissues would result in a lowered respiratory quotient. Whether such extremely low respiratory quotients can be sustained for long periods of time by seeds lying viable but dormant in the soil is a question. It should be emphasized that many variable factors which doubtless affect their life span in the soil were not present in the controlled experiments reported here. However, these studies indicate some of the fundamental changes which take place in seeds during moist storage and may serve as a beginning for further investigations.

Thornton and Denny,[16] working with gladiolus corms which had been held dormant in moist soil, determined that, in the phase when carbon

dioxide was low, oxygen consumption was high. The volume of oxygen absorbed was usually two or three times as great as the volume of carbon dioxide given off. This is essentially the same relationship as reported here for dormant *Amaranthus retroflexus* seeds.

Approximately a ten-fold reduction in respiration has taken place in seeds of *Amaranthus retroflexus* by the end of a year of moist storage (TABLE 1). It is doubtful whether this reduction is sufficient to maintain viability for as long a period as 40 years.[6] Again, it should be pointed out that the conditions imposed upon the moist seeds in the experiments reported here were not strictly comparable to those which exist in natural soil burial. Physiological behavior of the seeds in the two environments must vary similarly. At a temperature of 15° C., for example, respiration would no doubt be lower at all times than at 20° C. and the imbibed seeds would probably still remain dormant.

Denny[8] reported that the respiration of gladiolus corms, which had been held in the dormant condition from October 1937 to May 1939, by placing the bulbs soon after harvest in moist soil and storing them at room temperature, exhibited very low respiration, not more than, and probably less than 10 per cent of the rate of non-dormant corms. This reduction was of the same magnitude as that reported here. Ota[12] also found that the carbon dioxide eliminated in the first 60 hours, after the upper seeds of *Xanthium* were placed on wet absorbent cotton in a respiratory chamber, was ten times that given off at the twentieth day. The upper seed of *Xanthium* exhibits delayed germination.

Catalase and oxidase activity, as well as respiratory intensity, has been shown by Crocker and Harrington[5] to be reduced in Johnson grass seeds held in a germinator at 20° C. for a year. The mechanism concerned in the reduction of respiration in dormant imbibed seeds is not known. That the seed coat membranes are responsible for it is not likely, since it has been shown for *Cucurbita pepo*[4] that, after prolonged contact with water, the permeability of the seed coat to gases increased.

Additional studies on respiration and other processes taking place within the seed are needed for explanation of the physiologic responses of these seeds.

Impatiens balsamina. From the data in TABLE 2, we see a trend toward reduction in respiratory activity, as well as in respiratory quotient, which is in general agreement with results from *Amaranthus* seeds. These seeds were of the 1941 crop and tests earlier than 28 days of moist storage were without value, because the respirometer was not reliable. It should be noted that these measurements were for 100 seeds and for a 24-hour period.

TABLE 2

RESPIRATION OF SEEDS OF *IMPATIENS BALSAMINA* HELD ON MOIST GLASS WOOL AT 20° C. FOR DIFFERENT LENGTHS OF TIME AFTER HARVEST

(Carbon Dioxide Given Off and Oxygen Taken up Expressed as Ml. per 100 Seeds for a Period of 24 Hours)

Days in moist storage	CO_2	O_2	CO_2/O_2
28	0.46	1.12	0.41
56	0.22	0.40	0.55
77	0.24	0.50	0.48
119	0.08	0.70	0.11
365	0.12	0.54	0.22

Rumex obtusifolius. Respiration measurements were made on seeds of both 1941 and 1942 lots of these seeds. The values obtained were not consistent, because mold formed on the seeds and could not be controlled. Sterilization with calcium hypochlorite failed to remedy the situation. Consequently, the actual amounts of oxygen used and carbon dioxide given off varied a great deal from test to test and actually increased somewhat with lengthened storage, due, no doubt, to the increased amount of mold. In spite of this, however, there was a general decline in the respiratory quotient with increased storage period from approximately 0.89, at the beginning, to 0.42 after 632 days, and 0 36 after 923 days of moist storage at 30° C.

GERMINATION

From time to time in the course of the experiments described above, special germination tests were conducted to determine the viability and germination behavior of seeds held moist and ungerminated. These tests involved special treatments which will be described below.

Results and Discussion

Amaranthus retroflexus. Seeds of these plants when freshly harvested, show a degree of dormancy which varies from year to year and, with seeds collected in various localities, any one year. When any freshly-harvested seed lot is placed under moisture conditions which favor germination, the temperature determines the number of seedlings obtained. This is shown in TABLE 3 for the two 1942 seed collections. These two lots were similar in their germination response. The best temperature for germination was constant 35° C. where 86 per cent seedlings were obtained for lot A and 73 per cent for lot B, immediately after collection. Furthermore, the temperature requirement was quite

TABLE 3

EFFECT OF TEMPERATURE ON THE GERMINATION OF FRESHLY-HARVESTED AND MOIST-STORED SEEDS OF *AMARANTHUS RETROFLEXUS* 1942 CROP

Germ. Temp., °C.	Per cent germination after moist storage at 20° C. for days			
	Collection A		Collection B	
	0	623	0	613
15	0	0	0	1
20	2	0	0	0
25	18	84	8	91
30	16	86	11	90
35	86	91	73	90
10 to 20*	0	21	0	12
10 to 30*	0	83	2	90
15 to 30*	1	84	6	83
20 to 30*	1	85	3	88

* Daily alternation. Cultures left for eight hours at the higher temperature and for sixteen hours at the lower temperature each day.

specific at that time. Some germination occurred at constant temperatures of 25° and 30° C., but the percentages here did not exceed 18 in either crop. Only occasional seedlings appeared over the rather wide range of other temperatures tried.

If these seeds were stored dry after harvest, a gradual change in their capacity for germination took place, so that, after two or three months, they germinated over a wide range of temperatures including 20° C. If, however, they were placed on a moist medium at 20° C. immediately following harvest, this after-ripening did not proceed but rather primary dormancy was maintained or a secondary dormancy developed. This resulted in a different pattern of germination behavior. These seeds were not all of equal dormancy, for a few germinated immediately, and others germinated at intervals with no apparent cause. Careful records of these germinations have been kept for one lot of 1941 seeds and for two lots produced in 1942. They have all exhibited a periodicity in germination. This periodicity has been remarkably uniform for the three lots. Within a period of two months after the seeds of 1942 crop collection A had been moistened and placed at 20° C., 4 to 32 seedlings had appeared in each of the cultures of 2500 seeds each (TABLE 4). Very few additional seedlings were obtained up to eight months of moist storage. Between eight and ten months many seeds germinated. The number varied greatly in the different cultures but germination occurred in all. In 1942 crop collection A, for example, four cultures produced less than 100 seedlings during that period; nine cultures produced from 148 to 339 seedlings; six cultures from 404 to 620 seedlings; and six cultures from 1043 to 1764 seedlings. Some increases were evident up to the fourteenth month of moist storage, but, by the

TABLE 4

AMARANTHUS RETROFLEXUS, 1942 CROP, COLLECTION A
(*Germination after Various Periods on Moist Glass Wool at 20° C., 2500 Seeds in Each Lot*)

Lot No.	Number of germinations at end of each 2-month period											Total germination
	2	4	6	8	10	12	14	16	18	20	22	
1	14	1	2	0	95	18	13	0	1	115	208	467
2	10	7	0	7	148	63	34	9	1	94	259	632
3	10	2	1	3	228	41	12	6	7	73	115	498
4	8	4	0	0	242	171	14	1	0	95	168	703
5	11	1	0	1	272	29	5	1	0	66	235	621
6	11	1	1	3	273	265	18	0	0	22	223	817
7	12	0	0	3	404	36	9	0	0	58	181	703
8	7	3	0	4	430	73	19	0	1	59	500	1096
9	32	0	0	14	474	27	18	0	1	41	182	789
10	10	3	1	5	524	41	24	3	0	28	246	885
11	13	1	1	4	1360	39	7	0	0	4	210	1639
12	12	1	0	5	1562	163	11	0	0	19	53	1826
13	11	2	3	2	1668	65	16	2	3	13	61	1846
14	11	3	3	6	1764	35	6	1	0	18	48	1895
15	13	2	0	2	29	199	28	2	2	49	936	1262
16	14	1	0	1	56	33	5	3	3	71	648	835
17	13	2	2	2	74	69	34	5	1	75	491	768
18	12	0	3	1	339	71	18	0	1	120	510	1075
19	9	5	0	3	1178	120	21	1	6	15	389	1747
20	10	0	0	1	215	19	10	1	1	25	1740	2022
21	11	2	0	4	255	25	35	1	2	65	1903	2303
22	8	3	1	3	284	36	21	2	1	27	1869	2255
23	11	0	0	2	612	12	8	1	0	16	1430	2092
24	6	1	0	4	620	97	33	6	0	25	1331	2123
25	4	2	0	2	1043	141	10	0	2	11	1074	2289

sixteenth month, only occasional seedlings were found. Most of the remaining seeds continued dormant until between the eighteenth and twentieth month of storage, when a second germination pulse was noted. Here, again, seedling production occurred over a period of approximately two months. The seed lots which had many germinations after ten months of storage yielded fewer seedlings in the second germination period, as was to be expected. Similarly, those with fewer germinations during the first period produced more seedlings during the second period.

The temperature of the 20° C. room, where these seeds were kept so that they could be examined and respiration tests made without temperature effects, could not be maintained during the summer months. Consequently, all of the seeds were transferred from the room to a 20° C. constant temperature chamber on June 29, 1943. This necessitated their removal from the oven to the laboratory for examination and watering. Since sensitivity to temperature change is characteristic of

these seeds, it was thought that this may have brought about the increased germination. This seems unlikely, however, in view of the fact that the germination pulse did not come until July 26, in some cases, and was complete by August 2. The cultures were left in the 20° C. oven with weekly removal to the laboratory for examination until the first part of October with very few additional germinations. At this time, the cultures were replaced in the 20° C. room where the second germination pulse occurred after about 20 months. By the end of the 20-month storage period (early June, 1944), it was necessary to move the cultures again. At this time, however, a large room with a constant temperature of 20° C. was available. Lots 1 to 19 inclusive, TABLE 4, were transferred to this room so that it was not necessary to subject them to laboratory temperature for examination. Lots 20 to 25 inclusive were transferred to the constant temperature chamber, as in the summer of 1943, to determine more definitely the effect of the brief periods at laboratory temperatures. It will be seen that lots 20 to 25 all produced more seedlings, as represented by totals, than the other lots, but it was demonstrated conclusively that the second germination pulse did not depend on a change of temperature.

Under natural storage conditions in the soil, *Amaranthus retroflexus* seeds germinate before ten months. Bibbey[2] remarks that seedlings of this form appeared May 11 at Saskatoon, Saskatchewan, Canada, and Martin[11] listed these seeds among those that germinate late in spring in Iowa. He reported the same germination from *Amaranthus retroflexus* seeds held either wet or dry in refrigerators at 5° C. from before frost when they were harvested till early June when they were planted. If the temperature at time of planting was high, after-ripening would not have been a factor in this case.

Juby and Pheasant[10] used 25 sets of 100 seeds each of *Helianthemum guttatum* and found that they exhibited a constant intermittent germination. Only two germination maxima were involved, since germination was complete after the second one. Many other authors have observed that certain weed seeds are periodic in their germination, but very little definite experimentation has been done along this line. It seems, from the evidence at hand, that within any one seed lot varying degrees of dormancy, either primary or secondary, are found. This ensures the continuance of the species over a period of many years.

The germination capacity of the ungerminated seeds at various temperatures after 623 and 613 days of moist storage at 20° C. is shown in TABLE [illegible]. It will be noted that such seeds have become less specific in their temperature requirement for germination than when they were

freshly harvested. This would help to account for their germination under natural conditions. Their high germination capacity is in contrast to their low respiration as noted above.

In connection with temperature response, lots of seeds stored moist at 20° C. for approximately two years were removed and given 1, 2, 3, 4, 16, and 24 hours at 35° C., after which they were replaced at 20° C. Eighty-six, 75, and 90 per cent germination resulted after 4, 16, and 24 hours pretreatment respectively. The control lot, in this case, gave 15 per cent germination due, no doubt, to disturbance in transfer from the original seed lot to another Petri dish. The germination after two and three hours at 35° C. was only slightly less than that obtained after the four-hour period. One hour was not sufficient to induce much germination but produced more seedlings than the control.

Other experiments have been performed, using seeds stored moist at 20° C., to determine additional factors which might disturb the delicate dormancy equilibrium and bring about germination. Mechanical agitation of the soil in which the seeds are buried seems to bring about germination. We have found repeatedly that rubbing the seeds in the palm of the hand for 30 seconds was sufficient to induce approximately 50 per cent of them to germinate. Increasing the rubbing period to two minutes increased germination to 83 per cent in seeds held moist and dormant for one year. Similar results have been obtained with three crops stored two years. This suggests a coat restriction. Sulphuric acid treatment for two minutes was not as effective, but increased germination over that of the control.

Failure of certain *Helianthemum* seeds to germinate was due to an impermeable seed coat, according to Juby and Pheasant.[10] They attributed the intermittent germination to two factors: one, the physiological dimorphism of the seed ("hard" and "soft"), and the other, the wide range of variability of permeability of the seed coats.

The effect of drying the seeds for three hours, one day, and three days was noted. Such seeds returned to a moist medium at 20° C. gave 61, 59, and 73 per cent germination with the control at 0 per cent.

From these results, the difficulties involved in separating the several factors determining germination are apparent. Also the importance of a technique which would permit manipulation for respiration with a minimum of disturbance of the seeds is emphasized.

The seeds pictured in the respiration tube in PLATE 1 were removed from the 20° C. room and placed in the tube for photographing. They were out of the 20° C. room three and one-half hours, after which they were replaced. This was on July 13. On July 17, there were 1771

seedlings in this culture. In this case, it is impossible to determine whether mechanical disturbance (forceps were used to distribute the seeds evenly over the surface of the glass wool) or increased temperature brought about germination, since experiments have proved that either is effective. Tests have indicated that light does not favor germination, so could be eliminated as a factor.

Rumex obtusifolius. Although reliable respiration measurements could not be made on these seeds held ungerminated at 30° C. because of infection with mold, germination behavior was ascertained at intervals. Germination tests made on freshly-harvested seeds of both 1941 and 1942 crops indicated up to 98 per cent germination at a constant temperature of 15° C. or daily alternating temperatures of 10° to 20°, 10° to 30°, 15° to 30°, and 20° to 30° C. Slightly lower seedling production was obtained at 20° C. When the maintained temperature was as high as 25° C., germination was lowered to about 40 per cent, and no seedlings were produced at a constant temperature of 30° C.

Lots of 1000 seeds, each placed on moist glass wool at 30° C. immediately after harvest, showed continued primary dormancy or induced secondary dormancy, as was exhibited by *Amaranthus retroflexus* seeds at 20° C. Unlike *Amaranthus*, however, *Rumex* seeds showed no definite periodicity of germination under these conditions. There was some indication of increased germination after six and eight months of moist storage, but this effect was not marked. Rather, regular germination occurring throughout storage up to three years was characteristic. The number of seeds germinating was not large. In two years of storage, for example, representative lots had produced totals of 24, 41, 30, 74, and 101 seedlings out of 1000 seeds.

At intervals, these moist seeds were tested for germination capacity and for special treatments which might break their dormancy. Seeds of the 1941 crop still gave 97 per cent germination at a daily alternating temperature of 20° to 30° C. in June 1944. Special tests intended to bring about germination at the inhibiting temperature of 30° C. proved that the seed coat played an important rôle. Ungerminated year-old seeds germinated 96 and 90 per cent at 30° C., when the coats were broken or removed. Sulphuric acid treatment for two minutes induced 60 per cent germination. Seventy-four per cent of intact seeds transferred at 5° C. for three days germinated upon replacement at 30° C. A combination of sulphuric acid treatment and four days at 5° C. resulted in 93 per cent sprouting at 30° C. Some of these effects are shown in PLATE 3 A. It will also be seen (PLATE 3 B) that *Rumex*

seeds after-ripen in dry storage. Here, the germination of seeds of the same original lot held dry and moist at 30° C. are compared.

Rubbing *Rumex* seeds was without effect on germination. Also, these seeds were not as sensitive to mechanical disturbance as were those of *Amaranthus*.

SUMMARY

Physiological studies were made on imbibed seeds which remain without germinating and viable over long periods of time. The test material consisted of seeds of *Amaranthus retroflexus*, *Impatiens balsamina*, and *Rumex obtusifolius*

Gaseous exchange of *Amaranthus retroflexus* seeds, measured at intervals of from 0 to 901 days of moist storage at 20° C., showed at least a ten-fold reduction in respiration. The beginning of this reduction became apparent very early (after two days), and was definite after eight days in moist storage. Decreased respiration was also noted for *Impatiens balsamina* seeds held moist at 20° C. for 28 to 365 days With increased length of time in moist storage, the respiratory quotient decreased.

Seeds of *Amaranthus retroflexus*, held in moist storage, showed a periodicity in germination which was apparently independent of external conditions. This indicated varying degrees of the primary dormancy or the induced secondary dormancy of the original lot of seeds. Moist *Amaranthus* seeds, held without germination at 20° C., could be induced to germinate at that same temperature by rubbing, by drying for three hours to three days, or by exposure to 35° C. for 12 to 24 hours. Germination also proceeded immediately after removal to higher constant or alternating temperatures.

Moist seeds of *Rumex obtusifolius*, held without germination at 30° C., could be made to germinate at this same temperature by removal of the coats, treatment with concentrated sulphuric acid for two minutes, or exposure to 5° C. for four days. Upon removal from 30° C. to lower constant temperatures or daily alternating temperatures, germination proceeded without further treatment.

Some of the implications of these responses are discussed.

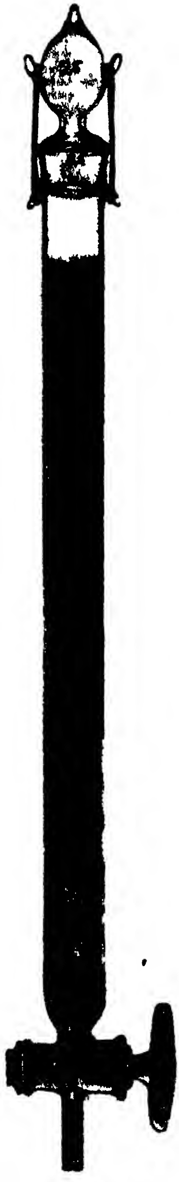
LITERATURE CITED

1. **Bailey, C. H., & A. M. Gurjar**
   1918. Respiration of stored wheat. Jour. Agric. Res. **12**: 685-713.
2. **Bibbey, R. O.**
   1935. The influence of environment upon the germination of weed seeds. Sci. Agric. **16**: 141-150.
3. **Brown, James W.**
   1939. Respiration of acorns as related to temperature and after-ripening. Plant Physiol. **14**: 621-645.
4. **Brown, R.**
   1940. An experimental study of the permeability to gases of the seed-coat membranes of *Cucurbita pepo*. Ann. Bot. **4**: 379-395.
5. **Crocker, William, & George T. Harrington**
   1918. Catalase and oxidase content of seeds in relation to their dormancy, age, vitality and respiration. Jour. Agric. Res. **15**: 137-174.
6. **Darlington, H. T.**
   1941. The sixty-year period for Dr. Beal's seed viability experiment. Amer. Jour. Bot. **28**: 271-273.
7. **Dennis, L. M., & M. L. Nichols**
   1929. Gas analysis. Rev. ed. 499 pp. Macmillan Co., New York.
8. **Denny, F. E.**
   1939. Respiration of gladiolus corms during prolonged dormancy. Contrib. Boyce Thompson Inst. **10**: 453-460.
9. **Haldane, J. S., & J. Ivon Graham**
   1935. Methods of air analysis. 4th ed. 176 pp. Charles Griffin & Company, Limited, London.
10. **Juby, D. V., & J. H. Pheasant**
    1933. On intermittent germination as illustrated by *Helianthemum guttatum* Miller. Jour Ecol. **21**: 442-451.
11. **Martin, John N.**
    1943. Germination studies of the seeds of some common weeds. Proc. Iowa Acad. Sci. **50**: 221-228.
12. **Ota, Junji**
    1925. Continuous respiration studies of dormant seeds of *Xanthium*. Bot. Gaz. **80**: 288-299.
13. **Sherman, Hope**
    1921. Respiration of dormant seeds. Bot. Gaz. **72**: 1-30.
14. **Stiles, Walter, & William Leach**
    1933. Researches on plant respiration. II. Variations in the respiratory quotient during germination of seeds with different food reserves. Proc. Roy. Soc. [Lond.] B. **113**: 405-428.
15. **Stoward, Frederick**
    1908. On endospermic respiration in certain seeds. Ann. Bot. **22**: 415-448.
16. **Thornton, Norwood C., & F. E. Denny**
    1941. Oxygen intake and carbon dioxide output of gladiolus corms after storage under conditions which prolong the rest period. Contrib. Boyce Thompson Inst. **11**: 421-430.
17. **Woo, M. L.**
    1919. Chemical constituents of *Amaranthus retroflexus*. Bot. Gaz. **68**: 313-344.

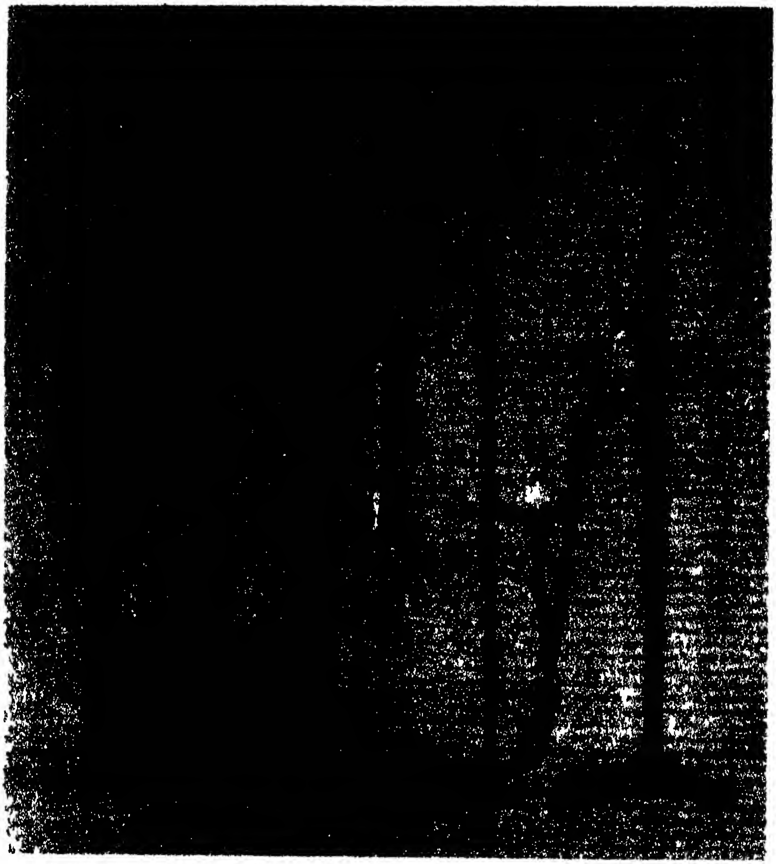
PLATES 1-3

PLATE 1

Seeds of *Amaranthus retroflexus*, 1942 crop, Collection B, on strip of moist glass wool on paraffined cheesecloth in a respiration tube Held moist at 20° C from October 13, 1942 to July 13, 1944



BARTON RESPIRATION AND GERMINATION OF SEEDS



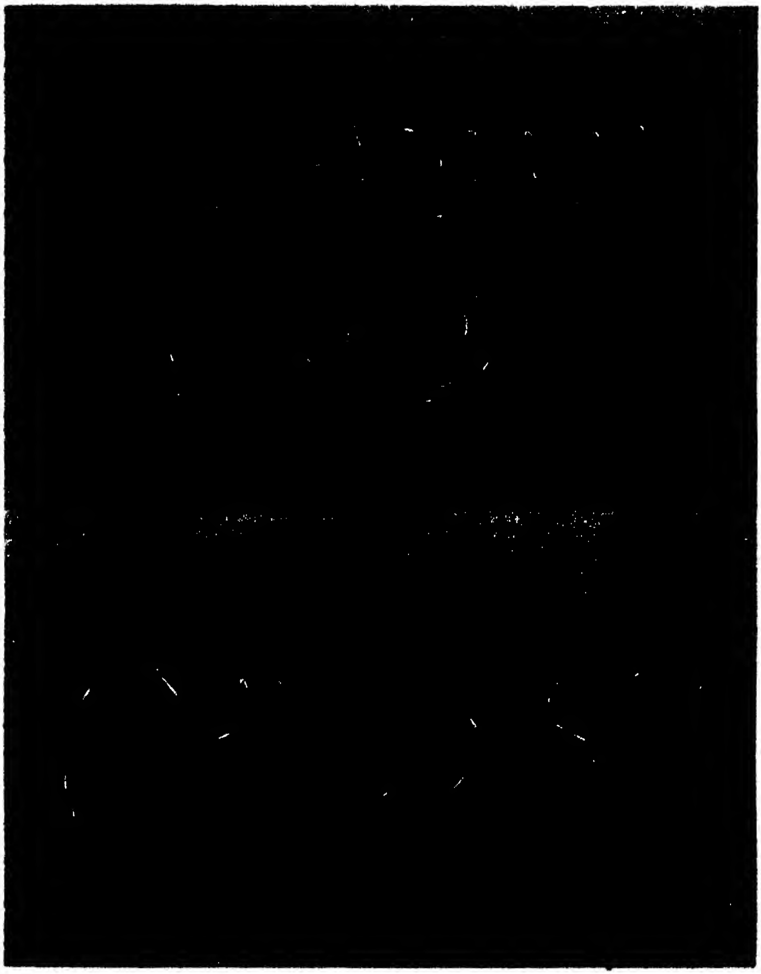
BARTON RESPIRATION AND GERMINATION OF SEEDS

PLATE 2

Modified respirometer for testing gaseous exchange of seeds

PLATE 3

Rumex obtusifolius, 1941 crop. Photographed December 9, 1942. A. Seeds held moist at 30° C. from November 18, 1941 to November 30, 1942. Top row: germination temperature 30° C. (0%). Middle row: germination temperature 15° to 30° C. daily alternation (96%). Bottom row: germination temperature 30° C. after treatment for two minutes with concentrated sulphuric acid, followed by three days on moist filter paper at 5° C. (93%). B. Germination temperature 30° C. Top row: seeds held moist at 30° C. from November 18, 1941 to November 30, 1942 (0%). Bottom row: seeds held dry at 30° C. from November 18, 1941 to November 30, 1942 (81%).



BARTON RESPIRATION AND GERMINATION OF SEEDS

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES

VOLUME XLVI, ART. 5. PAGES 209 TO 346

NOVEMBER 30, 1945

THE DIFFUSION OF ELECTROLYTES AND MACROMOLECULES IN SOLUTION*

BY

L. G. LONGSWORTH, CHARLES O. BECKMANN, MARGARET M. BENDER, EDWARD M. BEVILACQUA, ELLEN B. BEVILACQUA, DOUGLAS M. FRENCH, A. R. GORDON, HERBERT S. HARNED, LARS ONSAGER, JEROME L. ROSENBERG, AND J. W. WILLIAMS

CONTENTS

* This series of papers is the result of a conference on the Diffusion of Electrolytes and Macromolecules in Solution held by the Section of Physics and Chemistry of The New York Academy of Sciences, October 27 and 28, 1944.

Publication made possible through a grant from the Conference Publications Revolving Fund.

THE DIFFUSION OF ELECTROLYTES AND MACROMOLECULES IN SOLUTION: A HISTORICAL SURVEY

By L. G. Longsworth

From the Laboratories of The Rockefeller Institute for Medical Research, New York, N. Y.

INTRODUCTION

The general subject of molecular diffusion in solution was reviewed in 1934 by Williams and Cady[1] and in 1936 by Duclaux.[2] The diffusion of proteins was discussed recently by Neurath[3] and also by Edsall and Mehl.[4] In their adaptation of the theory of absolute reaction rates to the diffusion process Eyring and his associates[5, 6] have considered much of the available data on the diffusion of neutral molecules. The modern theory of the diffusion of electrolytes in solution has been presented by Harned and Owen.[7] With the exception of the work by Duclaux, all of these papers are readily available to American investigators. How, therefore, can a historical survey of diffusion be justified as an introduction to this monograph? The justification is twofold. Whereas each of the following papers has been prepared by specialists it has not been possible to cover all of the many aspects of the subject in this manner. In describing briefly some of the available methods that are not considered by the other conference participants, I have sough to overcome, partially at least, this defect. Secondly, this survey differs somewhat from those mentioned above in its emphasis on the experimental methods.

[1] **Williams, J. W., & L. C. Cady.** Chem. Rev. **14:** 171. 1934.

[2] **Duclaux, J.** "Diffusion dans les liquides." Actualités Scientifiques et Industrielles No. **349.** Hermann et Cie. Paris. 1936; "Diffusion dans les gels et les solides." Actualités Scientifiques et Industrielles No. **350.** Hermann et Cie. Paris. 1936.

[3] **Neurath, H.** Chem. Rev. **30:** 357. 1942.

[4] **Edsall, J. T., & J. W. Mehl.** Translational Diffusion of Amino Acids and Proteins, chap. 18 in **Cohn, E. J., & J. T. Edsall.** Proteins, amino acids and polypeptides. Reinhold Publishing Corp. New York. 1943.

[5] **Glasstone, S., K. J. Laidler, & H. Eyring.** The theory of rate processes. McGraw-Hill Book Co., Inc. New York. 1941.

[6] **Kincaid, J. F., H. Eyring, & A. E. Stearns.** Chem. Rev. **28:** 301. 1941.

[7] **Harned, H. S., & B. B. Owen.** The physical chemistry of electrolytic solutions Reinhold Publishing Corp. New York. 1943.

FICK'S LAWS OF DIFFUSION

Any history of diffusion in solution must begin with the theoretical work of Adolf Fick[8] and the experimental work of Thomas Graham.[9] Through his diffusion studies, Graham was led to the differentiation of solutes into colloids and crystalloids and to the discovery of dialysis as a means of separating the two. Fick also did some experimental work on diffusion but the confirmation of the laws bearing his name is due largely to the work of others. If the diffusion process is restricted to a vertical column of uniform cross-section, as is the almost universal practice, the quantity S, of solute diffusing through unit area in a given horizontal plane in unit time is proportional to the concentration gradient, $\partial c/\partial h$, at that level, *i.e.*

$$S = -D\frac{\partial c}{\partial h} \tag{1}$$

The proportionality factor, D, is the diffusion coefficient. If the flux, S, is expressed in grams per cm.$^{-2}$ per second, then the concentration, c, must be in grams per cm.^{3} Since S has the dimensions of m/l^2t, whereas those of $\partial c/\partial h$ are m/l^4, D has the dimensions of l^2/t and is usually expressed in cm.^{2} per second. Equation (1) is Fick's first law and contains the two dependent variables S and c. One of these, S, may be eliminated by noting that if the flux is given by Equation (1) for the level h, that at a neighboring level, $h + \Delta h$, will be $S + (\partial S/\partial h)\Delta h$. The accumulation of material in the layer between h and $h + \Delta h$ will thus be the difference between that entering, S, and that leaving, $S + (\partial S/\partial h)\Delta h$, and will amount, in the time, Δt, to $-(\partial S/\partial h)\Delta h \cdot \Delta t$. This difference, divided by the volume (Δh in this case since S refers to unit area) in which the change occurs, is the increment of concentration, Δc. In the limit, therefore,

$$\frac{\partial c}{\partial t} = -\frac{\partial S}{\partial h}$$

Elimination of S between this expression and Equation (1) then gives, if the diffusion coefficient is a constant, D_0, independent of the concentration

$$\frac{\partial c}{\partial t} = D_0\frac{\partial^2 c}{\partial h^2} \tag{2}$$

---

[8] Fick, A. Pogg. Ann. **94:** 59. 1855.
[9] Graham, T. Phil. Trans. **140:** 1, 805. 1850; **141:** 483. 1851, **144:** 177. 1854; **151:** 183. 1861.

This is Fick's second law. He pointed out the close analogy between diffusion and heat conduction and much of the mathematics of the latter process can be immediately adapted to diffusion problems. As will become apparent later in this survey, Fick's laws are limiting laws, strictly valid only at very low concentrations of the solute, a fact not fully appreciated for some time after their formulation.

Much of the theoretical work on diffusion has been concerned with the solution of Equation (2) for the various boundary conditions encountered in the experimental determination of diffusion coefficients These conditions fall into the following three classes:

BOUNDARY CONDITIONS IN DIFFUSION

(1) *Free diffusion.* The first is free diffusion from an initially sharp boundary between solution and solvent, or a more dilute solution, in a column in which the composition at the bottom and at the top of the column remains unchanged during the period of observation. The solution of Equation (2) for this case is the probability or error integral The following derivation is based on the requirement, first noted by Boltzmann,[10] that, although c varies with both h and t, these two variables must always occur in the ratio, h^2/t, for D_0 to have the proper dimensions. Consequently, if diffusion proceeds from an initially sharp boundary at $h = t = 0$ a new variable, y, may be defined by the relation $y = h/\sqrt{t}$. Then

$$\frac{\partial}{\partial t} = \frac{\partial y}{\partial t}\frac{d}{dy} = -\frac{1}{2}\frac{h}{t^{3/2}}\frac{d}{dy}$$

$$\frac{\partial}{\partial h} = \frac{\partial y}{\partial h}\frac{d}{dy} = \frac{1}{\sqrt{t}}\frac{d}{dy}$$

and Equation (2) becomes

$$\frac{y}{2}\frac{dc}{dy} = -D_0\frac{d^2c}{dy^2} \tag{3}$$

Setting $p = dc/dy$, Equation (3) yields, on integration,

$$-y^2/4 = D_0 \ln Ip$$

and, on returning to the original variables h and t,

$$\frac{\partial c}{\partial h} = \frac{1}{I\sqrt{t}}e^{-\frac{h^2}{4D_0t}} \tag{4}$$

The constant of integration, I, may be evaluated from the condition

[10] Boltzmann, L. Wied. Ann **53**: 959. 1894.

that $\int_{-\infty}^{+\infty} (\partial c/\partial h)dh$, the total area of the gradient curve, must equal, at all times, the initial concentration difference, Δc. Since each end of the column is at a physical infinity in the case of free diffusion and since the function (4) is symmetrical about the concentration axis,

$$\Delta c = \frac{1}{I\sqrt{t}} \int_{-\infty}^{\infty} e^{-\frac{h^2}{4D_0 t}}\, dh = \frac{2}{I\sqrt{t}} \int_{0}^{\infty} e^{-\frac{h^2}{4D_0 t}}\, dh$$

This integral has the value $\sqrt{\pi D_0 t}$ [11] and $I = 2\sqrt{D_0 \pi}/\Delta c$. Equation (4) then becomes

$$\frac{\partial c}{\partial h} = \frac{\Delta c}{2\sqrt{\pi D_0 t}}\, e^{-\frac{h^2}{4D_0 t}} \tag{5}$$

and

$$c = \frac{\Delta c}{2\sqrt{\pi D_0 t}} \int_{h}^{\infty} e^{-\frac{h^2}{4D_0 t}}\, dh \tag{6}$$

All of the recent work on the diffusion of proteins with the absorption, scale and schlieren methods is in the category of free diffusion. In the absorption method, the column, in which free diffusion is proceeding, is photographed with monochromatic light that is absorbed by the solute, but not by the solvent If the solute obeys Beer's law, and if the exposure does not exceed the linear portion of the characteristic curve for the photographic plate, a microphotometer tracing of the latter will then be a plot of equation (6).* The refractive index methods yield, on the other hand, gradient curves defined by Equation (5) at least in so far as the diffusion is ideal and the refractive index is a linear function of the concentration.

(2) *Restricted diffusion.* In restricted diffusion from an initially sharp boundary, the concentrations at one or both ends of the column *change* during the period of observation. Of course, in any column of finite height a diffusion process that is initially free will eventually become restricted when the concentrations at the ends of the column begin to change. This transition from free to restricted diffusion divides the process into what Thovert[12, 13] has called the first and second periods. Since solute can neither enter nor leave the system, the boun-

[11] Peirce, B. O. A short table of integrals Ginn and Co New York 1910 p. 62.

* Since the gradient refracts the light as well as absorbing it, the column must be illuminated in such a manner that all of the deflected rays enter the camera

[12] Thovert, J. Compt. rend. Acad. 133: 1197. 1901; 134: 594, 826. 1902, 135: 579 1902; 137: 1249. 1903; 138: 481. 1904; 150: 270. 1910.

[13] Thovert, J. Ann. chim. et phys. (7) 26: 366. 1902; (9) 2: 369. 1914

dary condition in restricted diffusion is that the flux, and hence, from Fick's first law, the concentration gradient, be zero for all times at the top and bottom of the column. The solution of Equation (2) for this case is a Fourier series whose usefulness depends upon the rapidity with which the series converges. Owing to this question of convergence, "a simple modification in experimental details will often save an enormous amount of labor in the mathematical work," a quotation from Mellor.[14] The work described elsewhere in this monograph by Harned and French[15] affords a splendid example of restricted diffusion in which the interpretation of the results has been greatly simplified by proper design of the cell.

Both Stefan[16] and Kawalki[17] have developed the theory of restricted diffusion. One important case is that in which three volumes of solvent are superimposed on one volume of solution in a vertical channel of uniform cross-section. Diffusion then proceeds for a time, t, after which the column is divided into four layers of equal volume and numbered, from the bottom upward, I to IV. These successive layers, the thickness of each being defined as $2\Delta h$, are removed and analyzed. If the total quantity of solute in the column is 100 on an arbitrary scale, the quantities in each of the four layers vary with the argument, $D_0 t/(\Delta h)^2$, as shown in FIGURE 1. For a given half-thickness, Δh, and diffusion coefficient, D_0, the abscissae are thus proportional to the time. The Stefan-Kawalki tables, on which FIGURE 1 is based, were used by many early workers. They were computed, however, on the assumption that the coefficient, D_0, is independent of the concentration. In many researches in which the tables have been used, this assumption has not been justified.

(3) *Steady state diffusion.* If the lower end of the diffusion column has access to a reservoir of solution at one concentration and the upper end to a reservoir of solution at a lower concentration, or pure solvent, a concentration distribution that does not change with time is eventually established in the column. In this steady state, the flux at each level has a constant value, S_1. If D is not a function of the concentration, the gradient of the latter is also constant, from Equation (1), and the concentration varies linearly with the height of the column.* The

[14] **Mellor, J. W.** Higher mathematics for students of chemistry and physics: 490. Longmans, Green and Co. London. 1912.
[15] **Harned, H. S., & D. M. French.** Ann. N. Y. Acad. Sci. 46: 267. 1945.
[16] **Stefan, J.** Sitzungsber. k. Akad. Wissensch. Wien 79 (2): 161. 1879.
[17] **Kawalki, W.** Ann. Physik 52: 166. 1894.

* This is strictly true only when the diffusion coefficient referred to a fixed mark on the column is constant. For a coefficient that is constant in a frame of reference moving with the solvent, the concentration will not vary linearly with the height when the steady state of diffusion has been established. The *frame of reference* is discussed later under that heading.

work of Clack[18, 19] and, as a first approximation, the porous diaphragm method of Northrop and Anson[20] are examples of steady state diffusion.

INTEGRAL AND DIFFERENTIAL DIFFUSION COEFFICIENTS

In general, however, the diffusion coefficient is not independent of the concentration and it becomes necessary, therefore, to distinguish between an average, *i.e.* integral, value over a finite concentration in-

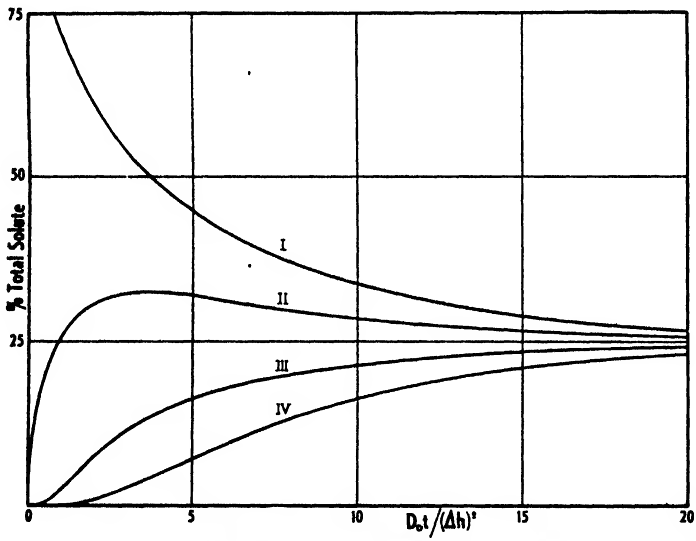


FIGURE 1. A plot of the Stefan-Kawalki tables for use in the study of diffusion by the method of layer analysis.

terval and the so-called differential value corresponding to a definite concentration. Clack's work, mentioned above as an example of steady state diffusion and described in more detail later in this paper, also illustrates the difference between these two types of coefficients. In his first experiments[18] he measured the flux, S_1, and computed an average value for the gradient from the concentration difference, Δc, between the reservoirs and the height, Δh, of the column. The value of the co-

[18] Clack, B. W. Proc. Phys. Soc. London **21**: 374. 1908; **24**: 40. 1912, **27**: 56. 1914, **28**: 49. 1916; **35**: 259. 1923.
[19] Clack, B. W. Proc. Phys. Soc. London **36**: 313 1924.
[20] Northrop, J. H., & M. L. Anson. Jour. Gen. Physiol. **12**: 543. 1929.

efficient given by the relation $\bar{D} = -S_1\varphi'(\Delta c/\Delta h)^{-1}$ is an integral value over the concentration interval, Δc. The factors, φ' in this equation and φ in Equation (7) below, correct for the counterflow of solvent as will be shown in the paragraph on *the frame of reference.* In later work, Clack[19] was able to measure not only S_1 but also the variation of the concentration with the height in the column and thus obtained a value of dc/dh for each value of the concentration. The diffusion coefficients computed from the relation

$$D = -S_1\varphi(dc/dh)^{-1} \tag{7}$$

are differential or true values. These are connected with the integral coefficients by the relation $\bar{D} = \frac{1}{\Delta c}\int_{c_1}^{c_2} D dc$. In general, the differential and integral values are the same only when the diffusion coefficient is a true constant. Elsewhere in this monograph, Gordon[21] shows how differential values may be computed from the integral coefficients in the more complex case, encountered in work with the porous diaphragm method, in which the concentrations in the reservoirs change during the experiment.

The full curves shown later in this paper in FIGURE 13 were computed from Clack's results and indicate the variation of the reciprocal of the concentration gradient, as abscissae, with the height for the steady state diffusion of saturated solutions of KCl, NaCl and KNO_3 into water. If the diffusion were ideal, these curves would be straight vertical lines corresponding to a constant gradient throughout the column. Quite obviously, they are not and the deviations therefrom indicate the manner in which the differential diffusion coefficient varies with the concentration.

BOLTZMANN'S RELATION

The differential diffusion coefficients have greater theoretical significance than the integral values. In principle, any of the available experimental methods will yield differential values if the concentration difference is made sufficiently small. As a matter of fact, if one allows a restricted diffusion process to proceed for a long time, a concentration difference that may have been quite large initially eventually becomes so small that the measured coefficient approaches the true or differential value for the mean concentration. For this reason, both Thovert[13]

[21] Gordon, A. R. Ann. N. Y. Acad. Sci. 46: 285. 1945.

and Lemonde[22] have stressed the advantages of working in the late stages of the second period of diffusion.

In many diffusion studies, however, relatively large differences of concentration have been used and the observations restricted to the early stages of the process. The significance of the coefficients obtained in these cases depends upon the experimental procedure and the method of computation and will be considered later in this survey in connection with the available procedures. Only in the steady state methods, where time disappears as a variable, can large concentration differences be used and differential values obtained directly with the aid of equation (1). In the free or restricted diffusion of a substance, the coefficient of which depends upon the concentration, Equation (2) must be rewritten in the form,

$$\frac{\partial c}{\partial t} = \frac{\partial}{\partial h}\left(D\frac{\partial c}{\partial h}\right)$$

Boltzmann[10] has examined this relation and suggested methods of computation based thereon. If c is a function only of $y(=h/\sqrt{t})$, then Equation (3) becomes

$$\frac{y}{2}\frac{dc}{dy} = -\frac{d}{dy}\left(D\frac{dc}{dy}\right)$$

This may be integrated and solved for D to give Boltzmann's relation

$$D = -\frac{1}{2}\frac{dy}{dc}\int_0^c y\,dc \tag{8}$$

The float method yields data that are readily interpreted with the aid of this equation and will be described later in this survey. The coefficient thus computed is, presumably, a differential value at the concentration c. Elsewhere in this monograph Beckmann and Rosenberg[23] have adapted Boltzmann's relation to the interpretation of data obtained with the refractive index methods.

THE THEORETICAL INTERPRETATION OF THE DIFFUSION COEFFICIENT

Although Long[24] suggested, in 1880, a relation between the diffusion coefficient of an electrolyte and the electric mobilities of its constituent

22 Lemonde, H. Ann. chim. et phys (11) 9: 539. 1938.
23 Beckmann, C. O., & J. L. Rosenberg. Ann N. Y. Acad Sci 46: 329 1945
24 Long, H. Ann. Physik. 9: 613. 1880

ions, it remained for Nernst[25] to formulate this relationship quantitatively. Nernst considered the driving force in diffusion to be the osmotic pressure gradient and this concept was retained by Schreiner[26] when he introduced activities into the theory. It was Hartley[27] who first suggested that the true driving force causing diffusion is the gradient of chemical potential of the diffusing substance and the derivation of the diffusion equation with the aid of this potential is certainly more direct, if no more rigorous, than with the osmotic pressure gradient.*

In the diffusion of a uni-univalent salt, for example, there is, in addition to the gradient of chemical potential, $\partial\mu/\partial h$, a gradient of electric potential, $\partial E/\partial h$, the so-called diffusion potential, due to the fact that the ions, having different mobilities in general, tend to diffuse at different rates. The velocity imparted to the i-th ion by these gradients is

$$v_i = -u_i\left(\frac{\partial\mu_i}{\partial h} \pm \frac{\partial E}{\partial h}\right) \tag{9}$$

in which u_i is the mobility. The minus sign in the parentheses is used if i is an anion. In the diffusion of a single binary salt, $\partial E/\partial h$ may be eliminated from the velocity equations for the two ion species since their velocities must be equal in order to preserve electrical neutrality. There results, if we recall that μ (in volts) $= \mu^\circ + (RT/F)$ $\ln fc$

$$S = cv = -\frac{2RT}{F}\frac{u_+u_-}{u_+ + u_-}\left(1 + \frac{d\ln f}{d\ln c}\right)\frac{\partial c}{\partial h}$$

Comparison with Fick's first law gives

$$D = \frac{2RT}{F}\frac{u_+u_-}{u_+ + u_-}\left(1 + \frac{d\ln f}{d\ln c}\right) \tag{10}$$

In this expression, R is the gas constant in volt · coulombs (joules) per degree per mole, 8.316, T the absolute temperature, F the faraday in coulombs per equivalent, 96500, and f the mean ion activity coefficient on a volume concentration scale, c. At infinite dilution, the values to be assigned to u_+ and u_- are the limiting mobilities obtained from conductance and transference data, *i.e.* $u_+ = t_+^\circ\, \Lambda^\circ/F$ and $u_- = (1 - t_+^\circ)$

[25] **Nernst, W.** Zeit. physik. Chem. 2: 613. 1888.
[26] **Schreiner, E.** Tidskr. Kem. o. Bergvaesen 2: 151. 1922.
[27] **Hartley, G. S.** Phil. Mag. 12: 473. 1931.

*Nernst's use of the osmotic pressure may account for the fact that he expressed the driving force in dynes and then reduced the ion mobilities to corresponding units. The evaluation of the constant factor entering into the theory can, however, be more readily understood if the mobilities retain their usual dimensions of $cm.^2 \cdot volt^{-1} \cdot sec.^{-1}$ and the chemical potential is expressed in volts. There is considerable justification for this procedure since e.m.f. measurements on galvanic cells are frequently used for the determination of the chemical potentials.

Λ^{0}/F. It is not permissible, however, to use these mobilities at finite concentrations.

The inter-ionic forces between the sodium and chloride ion moving, in a diffusion column at a given level of concentration, with the same velocity in one direction, are different from the forces between these ions when they are moving, in a homogeneous solution at the same concentration, in opposite directions under the influence of an external electric field. Hartley[27] showed, in 1931, that if the electric mobilities corresponding to finite concentrations were substituted in Equation (10), the values of D thus computed were too low. It remained, however, for Onsager and Fuoss[28] to extend the inter-ionic attraction theory to the diffusion problem with the result that, in the region of concentration for which the theory is valid, the diffusion mobilities for the ions of most salts are slightly greater than their limiting electric values and increase slowly with increasing concentration. Most of the change in the diffusion coefficients with the concentration is governed, however, by the thermodynamic factor $(1 + d \ln f/d \ln c)$. As Gordon[21] has emphasized elsewhere in this monograph, the only available experimental data of sufficient accuracy to permit a test of equation (10) are those of Harned and French[15] and the theory is confirmed within the limits of error of their data.

The Onsager-Fuoss theory promises to be of great value in the interpretation of experimental data and in the testing of experimental procedures. The limiting mobilities are known with high precision from conductance and transference measurements and the activity correction is also known quite accurately. Moreover, the success of Onsager's treatment of the variation of the electric mobility[29] with concentration inspires confidence in the corresponding treatment of the diffusion mobility. Any experimental procedure that can be adapted to salt solutions of 0.03 to 0.05 N, or less, and does not yield the theoretical value for the diffusion coefficient is to be regarded with suspicion.

The situation with respect to the diffusion of uncharged molecules is less satisfactory. Protein ions must be placed in this category, since the extrapolation of their electrophoretic mobilities to zero ionic strength as a means of determining the friction coefficient is not reliable and also involves uncertainty as to the charge on the protein. The interpretation by Einstein[30, 31] and Smoluchowski[32] of diffusion as a re-

[28] Onsager, L., & R. M. Fuoss. Jour. Phys. Chem. 36: 2689. 1932.
[29] Onsager, L. Physik. Zeit. 27: 388. 1926; 28: 277. 1927.
[30] Einstein, A. Ann Physik. 19: 371. 1906.
[31] Einstein, A. Zeit. Elektrochem. 14: 237. 1908.
[32] Smoluchowski, M. von. Ann Physik 21: 756. 1906.

sult of the thermal agitation of the diffusing molecules affords a detailed microscopic picture of the process. It does not, however, help the experimentalist who is testing his procedures, since sufficiently precise measurements of Brownian movement are not yet possible. The same is true of the Sutherland,[33] Stokes-Einstein[34] and Eyring[5] relations connecting the diffusion coefficient with the geometry of the diffusing molecule and the viscosity of the medium. As long as the diffusion measurements themselves remain the most reliable means for determining the mobility, or its reciprocal, the friction coefficient, there is no way of obtaining an independent check on the results. A careful investigator should, it seems to me, test his experimental procedures on dilute aqueous salt solutions before applying them to uncharged molecules

When more than two species of ion are present in the diffusion column, the individual ion velocities no longer need to have the same value in order to preserve electrical neutrality. The relations of the type of Equation (9) then become difficult to handle and no general solution has been obtained. The theory of such systems has been considered by Planck,[35] Taylor,[36] Guggenheim [37] and Sitte,[38] but the most complete development is given by Onsager in this monograph.[38a] Numerous experimental studies have been reported.[39-44] A limiting case of considerable practical importance is afforded by the diffusion of a large protein or colloidal ion in the presence of a neutral or buffer salt The equivalent weight of the protein ion is usually so great that a moderate concentration of salt throughout the diffusion column is sufficient to suppress the diffusion potential and thus allows the protein ion to diffuse independently of the other ions present.[45, 46]

EXPERIMENTAL METHODS

The method of layer analysis. It is now of interest to review the available experimental procedures in the light of the preceding survey of the theory. In much of the early work, the initial boundary was

[33] **Sutherland, W.** Phil. Mag. **1**: 781. 1903.
[34] **Einstein, A.** Ann. Physik **19**: 289. 1906.
[35] **Planck, M.** Ann Physik **39**: 161. 1890; **40**: 561. 1890.
[36] **Taylor, P. B.** Jour. Phys. Chem. **31**: 1478. 1927.
[37] **Guggenheim, E. A.** Jour. Am Chem. Soc. **52**: 1315. 1930.
[38] **Sitte, K.** Zeit. Physik **91**: 622. 1934.
[38a] **Onsager, L.** Ann. N. Y. Acad. Sci. **46**: 241. 1945.
[39] **Arrhenius, S.** Zeit. physik. Chem. **10**: 51. 1892.
[40] **Abegg, R., & E. Bose.** Zeit. physik. Chem. **30**: 545. 1899.
[41] **McBain, J. W., & C. R. Dawson.** Jour. Am. Chem Soc **56**: 52 1934
[42] **Sitte, K.** Zeit. Physik **91**: 642. 1934.
[43] **Teorell, T.** Jour Biol. Chem **113**: 735. 1936.
[44] **Vinograd, J. R., & J. W. McBain.** Jour. Am. Chem Soc. **63**: 2008 1941.
[45] **Hartley, G. S., & C. Robinson.** Proc Roy. Soc London, **A**, **134**: 20 1931.
[46] **Bruins, H. R.** Kolloid-Zeit **54**: 265. 1931: **57**: 152 1931· **59**: 263 1932.

formed, more or less undisturbed by mixing, by allowing the concentrated solution to flow under the dilute one. The apparatus of Scheffer[47] is typical, the dense solution being introduced through the tube, T, FIGURE 2. Diffusion then proceeded for a given time, usually until the concentrations at the ends of the column had changed, thereby restricting the process, after which successive layers were removed and

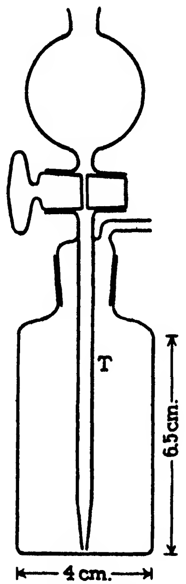


FIGURE 2 Scheffer's diffusion cell

analyzed. The diffusion coefficients were then computed with the aid of the Stefan-Kawalki tables, or similar ones corresponding to the number of layers analyzed in a particular experiment. The works of Arrhenius,[39] J. C. Graham,[48] Burrage,[49] the early investigations of Svedberg[50] and the great majority of Öholm's[51, 52] experiments were of this type.

---

[47] Scheffer, J. D. R. Zeit. physik. Chem. 2: 390. 1888.
[48] Graham, J. C. Zeit. physik. Chem. 50: 257. 1904.
[49] Burrage, L. J. Jour. Phys. Chem. 36: 2163. 1932.
[50] Svedberg, T., & A. Andreen-Svedberg. Zeit. physik. Chem. 76: 145. 1911.
[51] Öholm, L. W. Zeit. physik. Chem. 50: 309. 1905; 70: 378 1910
[52] Öholm, L. W. Meddelanden K. Vetenskapsakad. Nobelinstitut 2 (23): 1. 1913, (24): 1. 1913; (26): 1 1913.

A definite improvement in the results followed the introduction, by Schuhmeister,[53] of a shearing mechanism for forming a sharp boundary initially and for removing samples for analysis with a minimum of mixing. This device was refined by von Wogau[54] and by Dummer[55] and was used by Öholm[56] in his later work. An example of restricted diffusion with sheared boundaries is afforded by the work of Cohen and Bruins.[57] A diagram of their cell is shown in FIGURE 3. It consists of a pile of six glass discs, the thickness of each being 9.75 mm. The discs can be rotated independently about their common center. Holes, 2 cm. in diameter, drilled near the periphery through each of the four center plates form, when aligned as shown in FIGURE 3b, the diffusion column, the plates T and B serving as top and bottom, respectively. Initially, solution is introduced into the hole in plate I and solvent into those of the remaining three plates as shown in FIGURE 3a. Rotation of the plates until the holes are aligned, FIGURE 3b, then forms the initial boundary and diffusion proceeds for a given time. Each compartment is then isolated as shown in FIGURE 3c and its contents removed for analysis through small holes drilled for that purpose in the upper plates. Although not shown in FIGURE 3, provision is made for three simultaneous experiments in the one set of plates.

Cohen and Bruins made a careful study of the diffusion of 0.1 N potassium chloride into water. They allowed diffusion to proceed for 36 hours at 20°. The samples recovered from the cell were analyzed with the aid of an interferometer, the concentrations in the bottom compartment being about 0.035 N and that in the top compartment 0.015 N. As the mean of a large number of experiments, they compute, with the aid of the Stefan-Kawalki tables, the following values for the diffusion coefficient corresponding to each of the four compartments $D^*_{I} = 1.667$, $D^*_{II} = 1.685$, $D^*_{III} = 1.667$ and $D^*_{IV} = 1.678 \times 10^{-5}$. The average of these four values is 1.674×10^{-5} cm.^{2}/sec. and the average concentration is 0.025 N. The value given by the Onsager-Fuoss theory for 0.025 N is also 1.674×10^{-5}. This was computed on the basis of $\Lambda^\circ = 135.35$,[58] $t^\circ_+ = 0.4915$[59] and an ionic radius of 4.07.[60] The perfect agreement between the observed and computed values is

[53] **Schuhmeister, J.** Sitzungsber. k. Akad. Wissensch. Wien **79** (2): 603. 1879.
[54] **von Wogau, M.** Ann. Physik (4) **23**: 345. 1907.
[55] **Dummer, E.** Zeit. anorg. Chem. **109**: 31. 1919.
[56] **Öholm, L. W.** Meddelanden K. Vetenskapsakad. Nobelinstitut **2** (22): 1; (30): 1. 1913.
[57] **Cohen, E., & H. R. Bruins.** Zeit. physik. Chem. **103**: 337, 349, 404, 1923; **113**: 157. 1924.
r, **H. E., & A. R. Gordon.** Jour. Chem. Phys. **10**: 126. 1942.
, **R. W., D. J. LeRoy, & A. R. Gordon.** Jour. Chem. Phys. **8**: 418. 1940.
es, **D. A.** The principles of electrochemistry: 164. Reinhold Publishing Corp., New York. 1939.

somewhat ambiguous, however, since it will be noted that the values of D^* do not vary regularly with the concentrations in the four compartments.

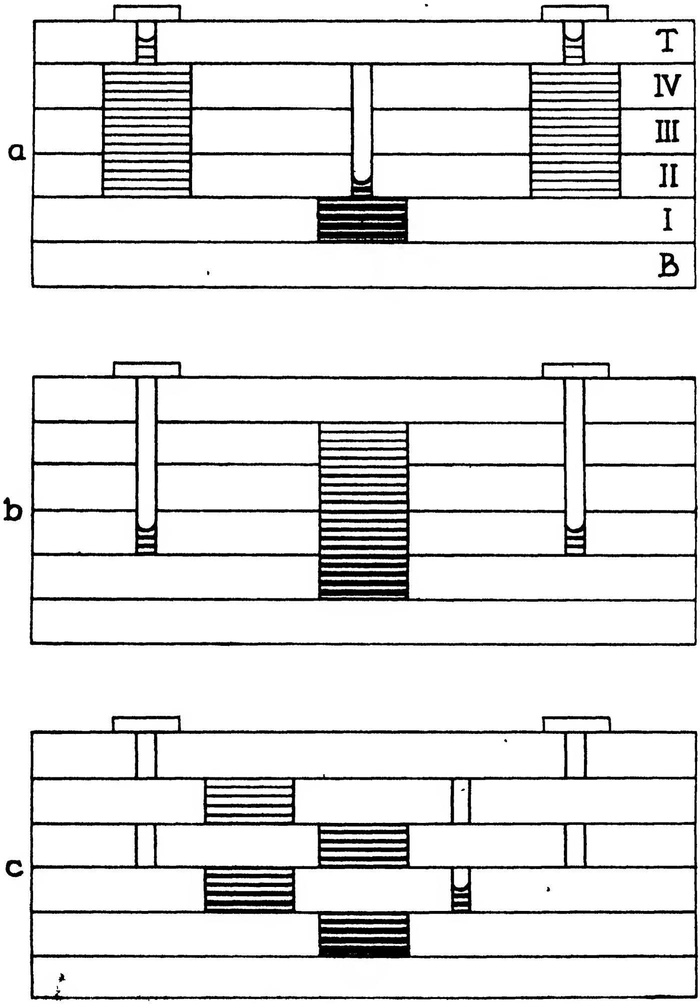


FIGURE 3. The sheared boundary cell of Cohen and Bruins

ments. This illustrates a fundamental difficulty in using a solution of the diffusion equation that is based on a constant coefficient for the in-

terpretation of an experiment in which that coefficient is a function of the concentration. The diffusion coefficient obtained in this case, D^*, is not a differential value, nor is it an integral value that can be related to either D or $\bar{D}$, as Hartley and Runnicles[61] have shown. It is unfortunate, indeed, that so many of the early investigators used the method of layer analysis, together with the Stefan-Kawalki tables, for materials whose diffusion coefficients are concentration dependent.

Of the more recent cells employing the shearing mechanism for the initial formation of a sharp boundary, and designed for use with the optical methods to be discussed later in this report, those of Neurath[62] and Tiselius[63] may be mentioned The sharpness of the original boundary formed with either of these cells leaves little to be desired. In the Neurath cell half of each window is, however, covered with a film of lubricant and the optical distortion arising therefrom has not been entirely eliminated. In the Tiselius cell, on the other hand, the boundary must be shifted[64] from behind the opaque horizontal plates of the cell and some slight mixing occurs during this displacement.

The float method. Fick,[8] Wilke and Strathmeyer[65] and Gerlach[66] have used the method of floats. This method is of interest since each float follows a level of constant concentration where the density of the solution is the same as that of the float. Moreover, if the initial boundary is at $h = t = 0$ the value of $h/\sqrt{t}$ for a given float remains constant. Graphic confirmation is thus afforded of the validity of Boltzmann's relation for those cases in which D depends on the concentration

In order not to disturb the diffusion, the floats should be small Gerlach used 2 to 3 cm. lengths of 1 mm. glass tubing. The ends are bent at right angles and sealed (FIGURE 4). The float thus takes up a horizontal position, that can be measured accurately, in the diffusion column. In a typical experiment on the free diffusion between methyl alcohol and nitrobenzene, 13 floats were used, 7 of which rose, as the diffusion proceeded, at different rates from the initial boundary, while 6 descended. The diffusion column was 6 cm. in diameter and 70 cm. in height and the period of observation was 14 days. The cross-section of each float was about 1 per cent of that of the column but no correction for its presence was made.

[61] **Hartley, G. S., & D. F. Runnicles.** Proc. Roy Soc London **A 168:** 401 1938
[62] **Neurath, H.** Science **93:** 431. 1941.
[63] **Tiselius, A.** Trans. Faraday Soc. **33:** 524. 1937.
[64] **Longsworth, L. G.** Ann N. Y Acad. Sci. **41:** 267. 1941
[65] **Wilke, E., & W. Strathmeyer.** Zeit. Physik **40:** 309 1926
[66] **Gerlach, B.** Ann. Physik. **10:** 437. 1931.

Gerlach obtained the values of $y(=h/\sqrt{t})$ corresponding to 13 known values of the concentration. He was thus able to prepare a plot of y versus c and from this to determine, by either analytical or graphical methods, the values of the coefficient, dy/dc, and the integral, $\int_0^c y\,dc$, appearing in Equation (8). Evaluation of the integral requires an extrapolation from the value of c for the lightest or heaviest float to pure solvent and it is in this region of concentration that D is usually changing most rapidly.

Optical methods. In the method of layer analysis the concentration distribution can be determined at only one time in each experiment, whereas the presence of the floats interferes somewhat with the diffusion process in that method. In order to avoid these limitations, several procedures have been developed for the determination of the concentration *in situ* without disturbing the diffusion. Optical methods

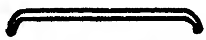


FIGURE 4 Gerlach's float

for accomplishing this have been the most popular, although conductivity[67, 68] and electromotive force measurements[69-72] have also been used in the case of salts. It was realized early in the last century that the gradient of refractive index accompanying a concentration gradient would act like a prism of variable power and deflect light rays accordingly. Gouy[73] was, however, one of the first to suggest the use of this property of the diffusion column for the determination of the concentration distribution. Although he did no quantitative experiments, his description of the propagation of light waves through such a column is very clear and is fundamental to all of the optical methods.

FIGURE 5 represents a vertical section through the essential portion of the optical system. An illuminated horizontal slit (not shown) to the left of the lens, L, is focused by that lens in the plane P. If the fluid in the cell C is pure solvent, say, the front of the wave emerging from the cell may be represented by the circular* arc *ab* and all of the light comes to focus at the level x_0 in the plane P. If, on the other

67 Haskell, B. Phys. Rev 27: 145 1908.
68 Lamm, O. Svensk Kem Tidskrift 51: 139, 1939, 55: 263 1943.
69 Weber, H. F. Wied. Ann 7: 469, 536. 1879.
70 Seitz, W. Ann. Physik 64: 759. 1898.
71 Procopiu, S. Ann. physique 9: 96. 1918.
72 Patterson, W. A., & J. T. Burt-Gerrans. Canadian Jour. Res. B 22: 5. 1944.
73 Gouy, G. L. Compt. rend. Acad. 90: 307. 1880.
* In actual practice the focal distances are usually so great in comparison with the height of the cell that the wave fronts are nearly plane.

hand, free diffusion is proceeding in the cell from an initially sharp boundary between solution and solvent at h_0, gradients of concentration will exist, after a time, in the region from h'' to h'. In FIGURE 5 the magnitude of the gradients has been indicated by the density of the shading. These gradients distort the wave front into the curve ah_0b'. The portion, $h''b'$, of the front emerging from the homogeneous solution below h'' is retarded, due to the higher refractive index of this material, relative to that, ah', emerging from the pure solvent, but both are still normal to x_0 and converge at that level. The normals to the

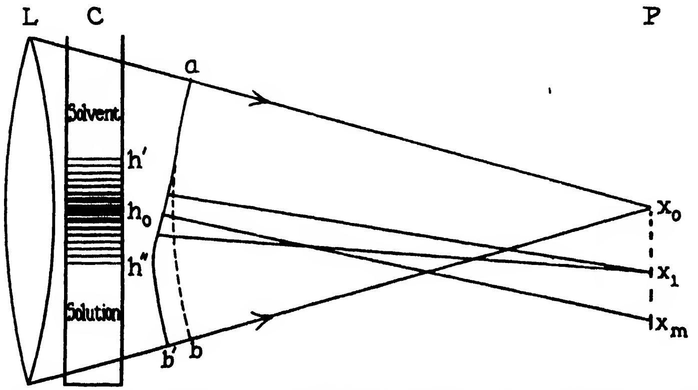


FIGURE 5. Diagram illustrating the interference phenomena accompanying the deflection of light by the gradients of refractive index in a freely diffusing boundary.

wave front in the transition region between h' and h'' intersect the plane P below x_0, however, and form, if the diffusion process is ideal, a system of alternate light and dark bands. These increase in width with increasing displacement from x_0 until a relatively broad diffuse band corresponding to the normal to the point of inflection at the level of maximum gradient, h_0, terminates the system at x_m. A photograph, taken by the author in the Rockefeller Institute Laboratories, of such a band system is shown in PLATE 1, A. In this instance, the light was not entirely monochromatic and may account for the fact, evident on close inspection of the photograph, that there is a superimposed periodicity in the system that causes the intensity and spacing of the bands to vary somewhat irregularly. The intensity at a given level, x_1 of FIGURE 5, depends upon the phase relations of the light from the two points on the wave front that are normal to x_1. If they are in phase a bright band results, otherwise a dark one. For ideal diffusion these points are equidistant from h_0, *i.e.* $h_0 + \Delta h_1$ and $h_0 - \Delta h_1$.

Consequently, if the cell is masked either above or below the level h_0, the light intensity in the plane P then varies monotonely with the height from x_0 to x_m. A quantitative treatment of this phenomenon, if one were available, would doubtless lead to greater precision in the use of the optical methods.

The first quantitative work on diffusion with the aid of an optical method is that of Wiener.[74] He also developed the relation between the concentration gradient and the light curvature. This development was made, however, in the terms of geometrical optics and does not account for such a phenomenon as that shown in PLATE 1, A. A diagram

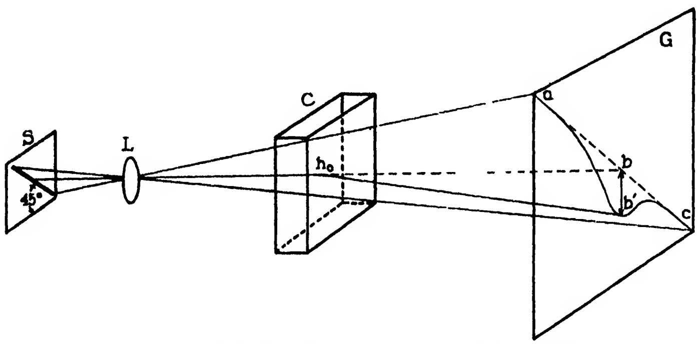


FIGURE 6. The optical system used by Wiener for the study of diffusion

of Wiener's apparatus is shown in FIGURE 6. An illuminated slit, S, inclined at 45° to the horizontal, is focussed by the lens, L, on a distant screen, G. The diffusion cell, C, is placed between the lens and the screen. If the solution in the cell is homogeneous the slit image at G is the straight line *abc*, also at 45° to the horizontal. If, on the other hand, free diffusion is proceeding in the cell from an initially sharp boundary in the horizontal plane at h_0, the pencils of light passing through the concentration gradients are deflected, generally downward, and the slit image is distorted as shown by the curve *ab'c* For ideal diffusion the curve *ab'c* is a plot of Equation (5) in coordinates whose axes make an angle of 45° with each other. The transposition to rectangular coordinates is straightforward. The precision obtainable with an optical method such as this depends very much upon the sharpness with which the distorted image is defined Heimbrodt[75] sought to improve the definition by restricting, with the aid of a narrow slit placed

[74] Wiener, O. Ann. Physik 49: 105. 1893
[75] Heimbrodt, F. Ann Physik (4) 13: 1028 1904

against the cell at an angle of 45°, the pencils of light forming the image of the slit S. Neither Wiener nor Heimbrodt published any photographs, however, and it is difficult to evaluate their work.

Thovert,[12, 13] in the course of diffusion studies extending over more than a decade, investigated several optical arrangements and published some interesting photographs. One such arrangement, also used by Lemonde,[22] was to measure, with the aid of a micrometer ocular, the maximum displacement, $x_m - x_0$ of FIGURE 5, as a function of the time. The arrangement of most interest to us, however, was one in which he obtained the gradient curve traced directly in rectangular coordinates

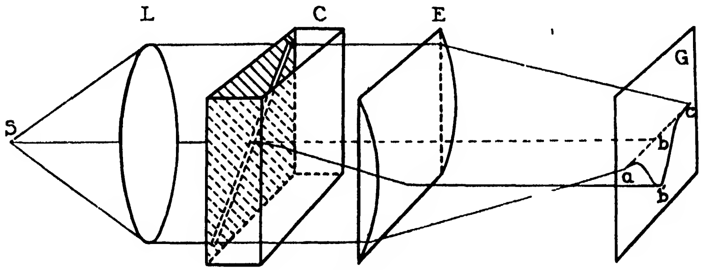


FIGURE 7. Thovert's cylindrical lens system

with the aid of a cylindrical lens as shown in FIGURE 7. Light from a point source, S, is collimated by the lens, L, and falls on the diffusion cell, C, whose front is masked by a slit inclined at 45°. The light emerging from the cell falls on the cylindrical lens, E, whose focal plane is at G and whose axis is horizontal. In the absence of a gradient in the cell, the light falling on the cylindrical lens is brought to focus in a straight horizontal line *abc*. If, however, gradients are present as in FIGURE 5, this line is distorted into the curve *ab′c*.

One of Thovert's photographs is reproduced in PLATE 1,B. The superimposed curves of this figure were obtained, as described above, at different stages in the free diffusion of 0.85 *N* sodium chloride into water. When allowance is made for the loss of detail in reproduction, these curves compare favorably with similar results obtained with the more modern optical methods. Thovert was keenly aware that the ultimate limitation of all optical methods is the wave nature of light. Anyone using such a method will derive inspiration from his papers.

With the procedures illustrated in FIGURES 6 and 7 the diffusion cell had to be as wide as it was tall. This disadvantage has been overcome in the more recent optical methods. These are the absorption

methods of Wiedeburg,[76] Svedberg[77, 78] and Quensel,[79] the scale methods of Littlewood[80] and Lamm,[81-87] the cylindrical lens method of Philpot[88] and Svensson[89] and the schlieren scanning method.[90] These have, however, been described so fully in conferences sponsored by this Academy[91, 92] and elsewhere that they will not be discussed further in this survey.

Micromethods. It is a consequence of the slowness of most diffusion processes in liquids that the experiments are usually of long duration. In addition to restricting the number of possible experiments, this also increases the probability that accidents, such as a thermostat failure, will intervene to spoil the determination. Since the concentration in a diffusion column is a function of $h/\sqrt{t}$ only, it is apparent that if the height of the column is decreased by a factor of 10, the time

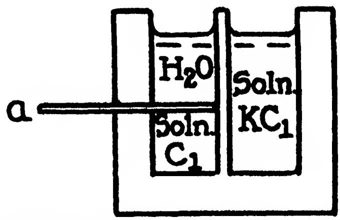


FIGURE 8. Furth's micro-colorimetric diffusion cell (2 × actual size).

required to reach a given stage in the process will be decreased by a factor of 100. The methods developed by Furth and his associates[93-97] have utilized this fact in order to reduce the duration of the experiments.

The cell used in the first method of Furth,[93] which is adapted to colored solutions only, is shown in FIGURE 8. It is placed on the platform of a microscope whose axis is horizontal. The microscope is provided

[76] Wiedeburg, O. Ann. Physik 41: 675. 1890.
[77] Svedberg, T., & H. Rinde. Jour. Am. Chem. Soc. 46: 2677. 1924.
[78] Tiselius, A., & D. Gross. Kolloid-Zeit. 66: 11. 1934.
[79] Quensel, O. A method of the determination of diffusion constants in "The Svedberg Memorial Volume": 198. 1944.
[80] Littlewood, T. H. Proc. Phys. Soc. London 34: 71. 1922.
[81] Lamm, O. Zeit. physik. Chem. 138A: 313. 1928; 143A: 177. 1929.
[82] Munter, E. Ann. Physik 11: 558. 1931.
[83] Franke, G. Ann. Physik 14: 675. 1932.
[84] Lamm, O., & A. Polson. Biochem. Jour. 30: 528. 1936.
[85] Lamm, O. Nova Acta Reg. Soc. Sci. Upsala IV, 10 (6). 1937.
[86] Polson, A. Kolloid-Zeit. 87: 149. 1939.
[87] Gralen, N. Kolloid-Zeit. 95: 188. 1941.
[88] Philpot, J. S. L. Nature 141: 283. 1938.
[89] Svensson, —. Kolloid-Zeit. 87: 181. 1939; 90: 141 1940
[90] Longsworth, L. G., & D. A. MacInnes. Jour. Am. Chem Soc. 62: 705. 1940.
[91] Bridgman, W. B., & J. W. Williams. Ann. N. Y. Acad. Sci. 43: 195. 1942.
[92] Longsworth, L. G. Ann. N. Y. Acad. Sci. 39: 187. 1939.
[93] Furth, R. Physik. Zeit. 26: 719. 1925.
[94] [illegible], R. Zeit. Physik 79: 280. 1932.
[95] Ullmann, E. Zeit. Physik 41: 301. 1927.
[96] Furth, R. et al. Kolloid-Zeit. 41: 300, 304. 1927.
[97] Furth, R. et al. Zeit. Physik 79: 275, 291, 306, 320. 1932; 91: 609, 617. 1934.

with a low power objective and with a micrometer ocular whose field is restricted to a narrow horizontal slit. The solution in the left hand side of the cell, FIGURE 8, is separated from an equal volume of water in the same side by means of a thin sliding partition, a. The same solution, after dilution by a known factor, k, is placed in the right hand side of the cell. Withdrawal of the partition, a, then forms the boundary. As diffusion proceeds, the horizontal slit in the ocular is shifted until the light intensity from the left hand side of the cell matches that

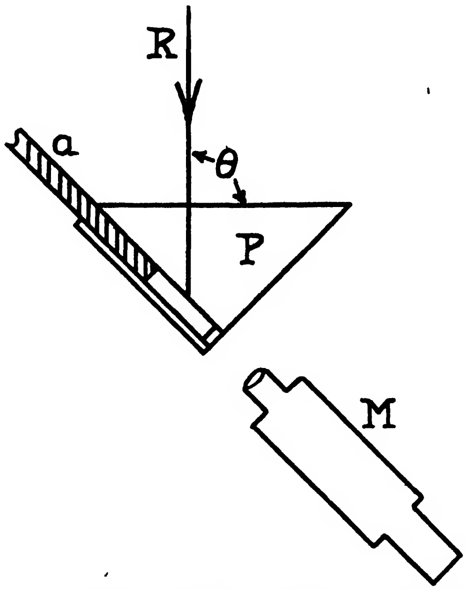


FIGURE 9. Zuber's micro-diffusion cell employing total reflection

from the right hand side. In this manner, the position of a layer of concentration kc_1 can be followed as a function of the time. Repetition of the experiment with a different value of k gives a similar curve for another concentration. With a cell of such small dimensions, it is practical to make several determinations in a day.

In order to extend the method to colorless solutions, Zuber[94] cemented the cell to a prism, P, as shown in FIGURE 9. This is a diagram of his arrangement as seen from above. The dimensions of the cell are similar to those of the left hand half of FIGURE 8, a sliding partition, a, being used in forming the boundary. The cell and prism are illuminated by

a collimated beam of monochromatic light. Whether or not a ray, R, gets through to the microscope, M, depends, however, upon the angle of incidence, θ, and the refractive index of the solution in the cell at the level at which the ray impinges. If the refractive index is sufficiently great, the ray, R, is totally reflected and does not enter the cell. The image of the cell in the microscope will thus consist, at each stage in the diffusion, of a dark and a light area, FIGURE 10. The illuminated region will be tangent to the side of the cell at some point, *h*, corresponding to the level in the cell at which the reflection is total. Since this corresponds to a definite concentration and can be followed with the ocular micrometer, one thus obtains, as with the microcolorimetric method, the position of a layer of constant composition as a function

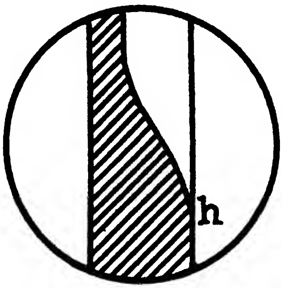


FIGURE 10. View of Zuber's cell as seen in the eyepiece of the microscope.

of the time. Moreover, by varying the angle of incidence, θ, a family of such curves may be obtained in one experiment. In a cell of these dimensions and for materials having diffusion coefficients of the order of 1×10^{-5}, the period of free diffusion is less than an hour and the experiments correspondingly short.

In the case of the micromethods, speed has been achieved, however, at a considerable sacrifice in precision. Ullmann[95] has mentioned 3 to 6 per cent as the probable error of a determination, but other results from Furth's laboratory[96, 97] indicate a somewhat larger error. Part of this is due to the mixing that occurs during the formation of the boundary. Such mixing is more serious in the experiments of short duration, characteristic of the micromethods, than in those extending over long periods of time. Moreover, the boundary formed by removal of a sliding partition is probably less sharp than one formed by a shearing mechanism.

Initial mixing and zero time. Guggenheim[37] has discussed the theory of diffusion for the case in which initial mixing occurs. Practically,

such mixing causes the concentration distribution at any instant to appear as though the boundary had been formed at some time before it actually was formed. If the diffusion process is otherwise normal, a quantitative estimate of the initial mixing may be obtained from the constant time increment that must be added to each time value in order to obtain a coefficient independent of that variable. Thus Lamm,[85] in an experiment to be described below and lasting about 10,000 seconds, had to add 75 seconds to each time reading in order to secure a constant coefficient for 0.1 N potassium chloride diffusing into water at 20°. No systematic study of the effect of initial mixing has been reported, however.

Precision of the optical methods. It is improbable that any of the recent optical methods have been used with more care than that exercised by Lamm in the experiment mentioned above. Using the scale method in conjunction with a cell of 6 cm. height, in which the boundary was formed initially with the aid of a removable partition, he obtained the value 1.667×10^{-5}. This was computed[87] from the relation $2D^{*}t = \sigma^2$, in which σ is the standard deviation of the gradient curve, and thus cannot be identified with either a differential or an integral coefficient. The theoretical concentration corresponding to a differential value of 1.667×10^{-5} is 0.03 N. This is somewhat less than the mean concentration, 0.05 N, of Lamm's experiment, but indicates substantial agreement with the theory. Since the coefficient for potassium chloride decreases some 10 per cent in the interval from 0 to 0.1 N, the concentration gradient curve for free diffusion over that range should be slightly skew on the dilute side. Actually, Lamm's curve was normal within his limit of error. This probably indicates that the optical methods in current use for the study of free diffusion, while capable of yielding results that are accurate within a very few per cent, are not yet sufficiently sensitive for work of high precision, especially in the region of dilute solutions for which the theory is valid.

The steady state method of Clack. Some of the difficulties that beset studies of either free or restricted diffusion vanish when one uses a steady state method. An example is afforded by the work of Clack.[18, 19] A diagram of his first apparatus is shown in FIGURE 11. Two bulbs, A and B, are connected by the tubes t and t′. These bulbs are filled, through the tube D, with the solution to be studied, connected with a fine wire, W, to the arm of a balance and immersed in a bath containing a large volume of water, or a dilute solution of the same salt as that in the bulbs. The salt present in the tube D diffuses upward into the water of the bath and the dilute solution thus formed just above the

upper end falls, under the influence of gravity, toward the bottom of the bath. In this manner, the salt concentration at the upper end of

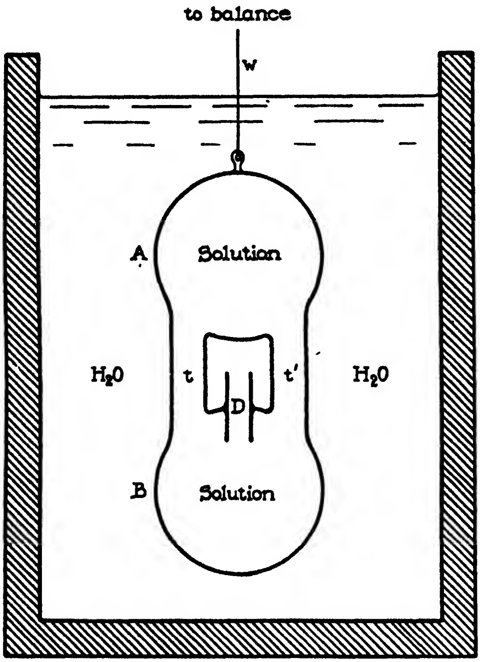


FIGURE 11. Clack's first apparatus for the study of steady state diffusion.

the tube is held at a constant value. Similarly, the salt entering the bottom of the tube D causes the solution just below the end of the tube to become diluted and it is replaced, therefore, by solution of the original concentration from the upper bulb A. After some time, depending on the temperature, the salt and the length of the tube D, a steady state of diffusion is established in that tube and the bulb system henceforth loses weight at a constant rate. From the value of this weight loss per unit time, the dimensions of the tube D and the density of the solution, an integral diffusion coefficient for the concentration interval between the top and bottom of the tube may be computed.

The time required to establish the steady state, 12 to 14 days in most of Clack's experiments, can be reduced by using a short tube. He observed, however, that for a given diameter of the tube, D of

FIGURE 11, the rate of weight loss was inversely proportional to its length only for tubes above a certain length. This critical length could be reduced by decreasing the tube diameter but the method then became insensitive due to the lower rate of weight loss. His solution of this difficulty was to use, instead of a single tube, a battery of short tubes of small diameter and we see here a step toward the porous diaphragm method.

In a later paper, Clack[19] describes a modified cell in which the steady state of diffusion is established in a channel of rectangular cross-section suitable for use with optical methods. He was thus able to determine the refractive index gradient at different levels in the column and hence differential diffusion coefficients as a function of the concentration. The optical arrangement is similar to one used by Thovert and is shown in FIGURE 12. The image of the horizontal slit, S, illuminated with monochromatic light, is formed in the plane of the ocular, E, by the lenses L_1 and L_2. The channel, C, in which the steady state of diffusion has been established, is masked by the screen, B, containing a compound slit, Y. In Clack's apparatus, this consisted of a pair of horizontal slits 0.2 mm. wide and 1.3 mm. apart. This compound slit produces an interference pattern in the ocular whose central band is more sharply defined than in the case of a simple slit. The screen B and the ocular E may be displaced vertically and for each position, h, of the slit Y there is a corresponding displacement, δ, of the central

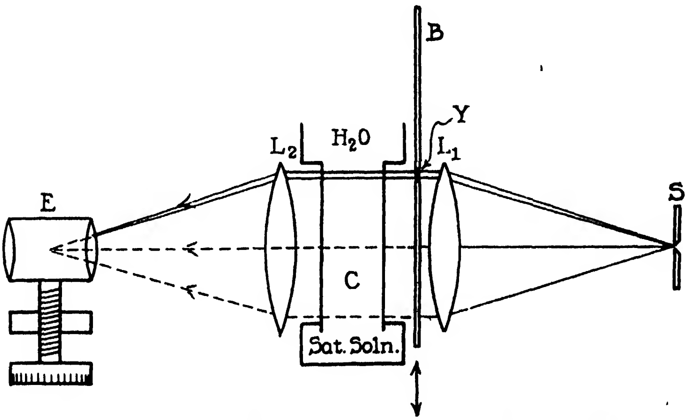


FIGURE 12. The optical system used by Clack in the determination of differential diffusion coefficients.

band in the ocular E from its position when the solution in the channel is homogeneous The displacement, δ, is given[74] by the product, $ab(dn/dh)$, in which a is the horizontal dimension of the column in the direction taken by the light, b is the distance from the cell to the ocular and (dn/dh) is the value of the refractive index gradient at the level h If n is known as a function of the concentration, both c and dc/dh can

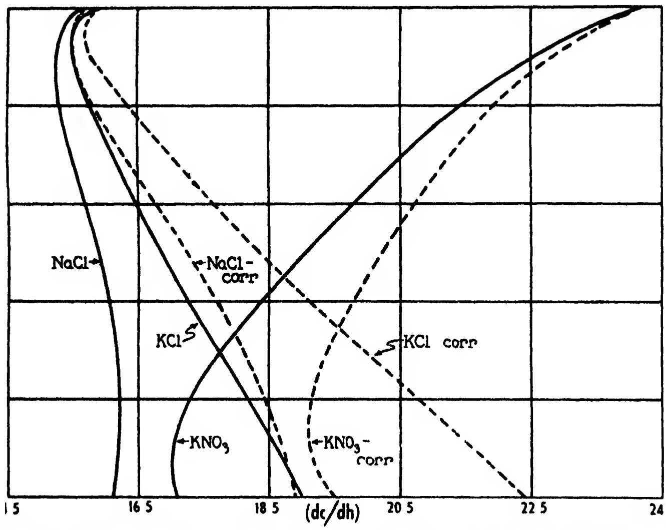


FIGURE 13 Clack's results for the steady state diffusion, at 18 5°, of NaCl, KCl and KNO_3 The reciprocal concentration gradients $(dc/dh)^{-1}$, are plotted as abscissae against the height in the diffusion column as ordinate The salt concentration at the top of the column $h = 0$ is essentially zero, while the bottom, $h = -5$ cm, is in contact with saturated solution The full curves represent the observed values of $(dc/dh)^{-1}$ The dashed curves correspond to values that have been corrected for solvent counterflow

be computed for each level in the column. Clack's values of $(dc/dh)^{-1}$ for the diffusion of saturated solutions of KCl, NaCl and KNO_3 into water* at 18.5° are represented by the full curves in FIGURE 13. Using the same method, Davies[98] has obtained data for $MgSO_4$ and $CdSO_4$.

The values of S_1 and φ in the relation

$$D = -S_1 \varphi (dc/dh)^{-1} \tag{7}$$

must also be known in order to compute absolute values of the differential diffusion coefficients. It will be recalled that S_1 is the observed

* Actually the upper reservoir, into which the salt diffused, contained a very dilute solution instead of pure water. Clack made a small correction for this.

[98] Davies, R. J. Phil. Mag. (7) 15: 489. 1933.

flux of salt in the steady state and φ is a factor correcting for the solvent counterflow. In Clack's improved cell, the lower reservoir is filled with a saturated solution in equilibrium with solid salt and can gain or lose matter only through the diffusion column, as in the cell shown in FIGURE 11. After the establishment of the steady state, this reservoir loses S_1 grams of salt, but gains S_0 grams of water, per unit time and unit cross-section of the column. The observed net loss, w, per unit cross-section is thus $S_1 - S_0$. Simultaneously, w' grams of salt in the lower reservoir dissolve and their volume, w'/d' (if d' is the density of the salt) occupied by saturated solution. If c'' and d'' are the concentration and density, respectively, of this solution, the volume w'/d' will contain $w'c''/d'$ grams of salt and $(w'd''/d' - w'c''/d')$ grams of water. Conservation of salt for the system then gives

$$w' - w'c''/d' = S_1$$

and of water

$$w'd''/d' - w'c''/d' = S_0$$

Elimination of w' between these two expressions gives

$$S_0/S_1 = (d'' - c'')/(d' - c'') = \rho$$

and

$$S_1 = w + S_0 = w/(1 - \rho)$$

in which $\rho = S_0/S_1$. With the aid of these relations, S_1 and ρ may be computed from the observed weight loss w and the known properties of the solutions involved.

The frame of reference The value of the ratio, $-S_1/(dc/dh)$, at each level in FIGURE 13 is a differential coefficient corresponding to the concentration at that level. It refers, however, to diffusion relative to a frame of reference fixed with respect to the apparatus whereas the values required are those relative to the solvent.* As mentioned above, the volume change in the lower reservoir accompanying the upward diffusion of salt is balanced by a downward movement of water. The salt thus diffuses against a countercurrent of solvent. Moreover, the velocity, v_0, of this countercurrent is not constant throughout the column, but, due to the volume occupied by the salt, is accelerated in the lower concentrated layers. Of course, the downward flux of water, S_0, remains constant at each level when the system is in a steady state.

The observed flux of salt, S_1, is the product, cv, of the concentration

* The use of the solvent as the stationary element appears to be well adapted to the interpretation of diffusion data obtained by a steady state method. In the following paper of this monograph Onsager shows, however, that other frames of reference can be used to advantage with other methods.

at a given level by the velocity, relative to the apparatus, with which the salt moves. The flux relative to the water will thus be

$$c(v + v_0) = S_1 + cv_0 = -D(dc/dh) \qquad (11)$$

If v_{0T} is the velocity with which water enters the top of the column where $c = 0$, its velocity at a level where the concentration is c will be $v_{0T}/(d - c)$, in which d is the density of the solution whose concentration is c grams per cc. Since S_0 and v_{0T} are almost identical for aqueous solutions, $v_0 = v_{0T}/(d - c) = S_0/(d - c) = \rho S_1/(d - c)$, and Equation (11) becomes

$$D = - S_1 \left(1 + \frac{c\rho}{d - c}\right) (dc/dh)^{-}$$

This is the relation obtained by Clack. The correction factor, φ of Equation (7), is thus seen to be $1 + c\rho/(d - c)$. The value of the differential diffusion coefficient referred to the solvent, D, is identical, at sufficiently low concentrations, with that referred to the apparatus, but becomes progressively larger as the concentration is increased. This is shown in FIGURE 13, where the dashed curves represent the reciprocal gradients after correction for the solvent counterflow, and are, therefore, proportional to D, the factor of proportionality being the flux of salt, S_1.

Clack's work has been discussed at some length because his results were found by Onsager and Fuoss to be most nearly in accord with the theory of the process. The data obtained with the aid of a steady state method are also well adapted to illustrate (1) the difference between integral and differential diffusion coefficients and (2) the importance of the frame of reference. The failure of many investigators to state what kind of a coefficient they are reporting makes it difficult to compare results on the same system when they are obtained by different methods or at different stages in the process. Finally, Clack's work is of particular interest to those of us who are interested primarily in the optical methods. Unlike the gradients of refractive index that arise in free diffusion, those present in a column in which a steady state has been established change relatively slowly with the height. Thus dn^2/dh^2 is small and many of the aberrations disappear that restrict the precision of the optical methods in the case of free diffusion. The difficulties attendant upon the development and maintenance of the steady state, although formidable, can be overcome and the reward, in Clack's case at least, was an outstanding contribution to our knowledge of diffusion.

The author realizes that, in concluding this survey now, he is omitting reference to many notable researches on diffusion. This is especially true of the work on materials of high molecular weight. This aspect of the problem is, however, treated comprehensively and critically in the accompanying papers by Bevilacqua, Bevilacqua, Bender and Williams[99] and by Beckmann and Rosenberg.[28] Finally, the porous diaphragm method has not been considered, since it is also discussed elsewhere in this monograph.[21]

ACKNOWLEDGMENT

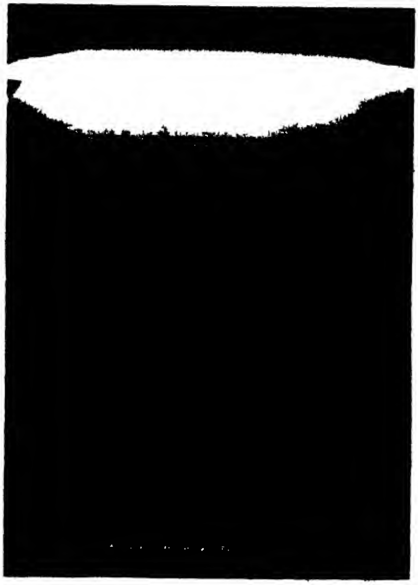
It is a pleasure to acknowledge my indebtedness to Dr. D. A. MacInnes, of these Laboratories, and to Dr. Lars Onsager, of Yale University, for their kindness in reviewing this survey prior to its publication

[99] **Bevilacqua, E. M., E. B. Bevilacqua, M. M. Bender, & J. W. Williams.** Ann. N Y Acad. Sci. **46**: 309. 1945.

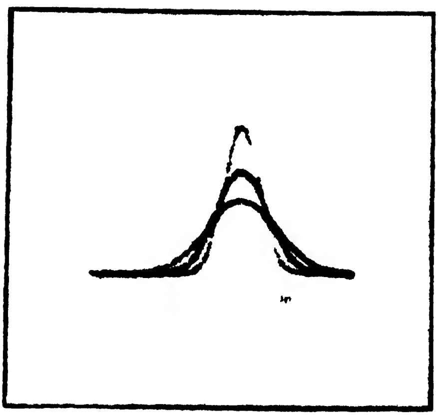
PLATE 1

A. Photograph of the interference pattern formed in the focal plane of the light after deflection by a free diffusion gradient.

B. The curves of refractive index gradient at different stages in the free diffusion of 0.85 *N* NaCl into water. These were photographed directly by Thovert with the optical system shown in Figure 7.



A



B

LONGSWORTH DIFFUSION OF ELECTROLYTES

THEORIES AND PROBLEMS OF LIQUID DIFFUSION

By Lars Onsager

From the Sterling Chemistry Laboratory, Yale University, New Haven, Connecticut

INTRODUCTION

The theory of liquid diffusion is relatively undeveloped Most of the theoretical problems involved are complicated and interesting simple results of fundamental significance do not come easily. Further to discourage the theorist, the basis for an inductive approach is neither broad nor firm.

Various theoretical considerations lead us to expect a fairly close correlation between diffusion and viscous flow, and the limited body of evidence tends to bear this out. Since viscosities are much more easily measured and we know more about them, the correlation is helpful in various ways. From measurements of viscosity, we can estimate coefficients of diffusion as to order of magnitude. Moreover, generalizations derived from our experience with one phenomenon may be extended tentatively to the other. Where concentrated solutions are concerned, their colligative properties are also involved.

The diffusion of highly ionized electrolytes is intimately related to the migration of the ions in an electric field. It is proper and convenient to deal with the two processes as different aspects of the same general phenomenon. Ions move in either case and, as a matter of fact, diffusion alone can cause electric currents.

Even at quite low concentrations, the properties of electrolytes are complicated by various effects which are now recognized as consequences of the Coulomb forces between the ions. The quantitative theoretical prediction of these effects has greatly advanced our understanding of electrolytes. The properties of concentrated electrolytes, however, cannot be predicted from general theory any better than those of other concentrated solutions.

Very little attention has been given to quantitative studies of diffusion in systems of more than two components, although this general case of diffusion is inevitably involved in many techniques of electrochemistry on every scale from micrograms to tons. One can understand how the experimental difficulties have deterred investigators However, a long period of inactivity has probably added an element of mental inhibition From sheer weight of tradition the conclusions of

relatively primitive theories tend to be accorded some of the reverence that is properly given to practical knowledge.

It is a striking symptom of the common ignorance in this field that not one of the phenomenological schemes which are fit to describe the general case of diffusion is widely known.

SYSTEMS OF DESCRIPTION

The Generalization of Fick's Law

Diffusion in a liquid system of two components is governed by Fick's law[1]

$$\mathbf{J}_i = -D\nabla c_i; \qquad (i = 1,2) \tag{1}$$

where c_1, c_2 denote the concentrations per unit volume, and $\mathbf{J}_1$, $\mathbf{J}_2$ the flows in corresponding units. Equation (1) only describes the *relative* motion of the two components. Additional terms,

$$\mathbf{J}_i' = c_i\mathbf{v} \tag{2}$$

where the velocity $\mathbf{v}$ obeys the laws of hydrodynamics, will, in general, be superimposed. However, provided only that the volume changes due to mixing may be neglected, it is possible to arrange conditions such that $\mathbf{v} = 0$ everywhere. It is customary to take such an arrangement for granted.

For a system of $s > 2$ components we may assume, as a natural generalization of Fick's law, that the relative velocities of the different components will be linear functions of the concentration gradients. Thus

$$\mathbf{J}_i = -\sum_k D_{ik}\nabla c_k \tag{3}$$

apart from the hydrodynamic flow (2). We may require that diffusion shall cause no bulk motion of the liquid. As in the case of Equation (1), this condition amounts to nothing more than a convenient definition. In terms of the partial volumes $\bar{V}_1, \ldots, \bar{V}_s$ the bulk displacement of liquid is

$$\mathbf{v} = \sum_i \bar{V}_i \mathbf{J}_i \tag{4}$$

We therefore impose the conditions

$$\sum_i \bar{V}_i D_{ik} = 0 \tag{3a}$$

upon the coefficients of Equation (3). Furthermore, since the concen-

[1] Fick, A. Pogg. Ann. 94: 59. 1855.

tration gradients at constant pressure are subject to the condition

$$K\nabla P \equiv \sum_i \bar{V}_i \nabla c_i = 0 \tag{5}$$

(K = compressibility), we may impose another set of conventional conditions. The set

$$\sum_k D_{ik} c_k = 0 \tag{3b}$$

is to be preferred because it is easy to apply and it leads to a particularly simple form for the reciprocal relations which apply to diffusion.

The sets of Equations (3a) and (3b) contain altogether (2s-1) restrictions, for both imply

$$\sum_{i,k} \bar{V}_i D_{ik} c_k = 0$$

but otherwise the equations are independent. Accordingly there are $(s\text{-}1)^2$ independent coefficients of diffusion in a system of s components.

For the case $s = 2$ we obtain Equation (1) in disguise, with

$$D = D_{11}/\bar{V}_2 c_2 = -D_{12}/\bar{V}_2 c_1 = -D_{21}/\bar{V}_1 c_2 = D_{22}/\bar{V}_1 c_1 \tag{6}$$

It seems impossible to devise a simple general scheme of description which reduces directly to Equation (1) for the case $s = 2$. We may try to satisfy Equation (4) identically by some set of conditions other than Equation (3a). Thanks to Equation (5) this is possible; but we have only one arbitrary constant at our disposal, not enough to be of any help except for $s = 2$.

Reciprocal Relations, Dissipation-Function

From the hypothesis of microscopic reversibility[2] one may derive certain symmetry conditions which are presumably satisfied by the coefficients D_{ik} of Equation (3). While the conditions in question have never been put to a test, analogous relations which apply to thermoelectric phenomena,[3] electrolytic cells with liquid junctions[4, 5] etc.,[2] have been verified by numerous experiments, so that we have good reason for confidence in theoretical relations of this general type.

It has been shown[2] that when these reciprocal relations hold, the equations of transport processes can be expressed in terms of thermodynamic functions and the so-called *dissipation-function.*

[2] **Onsager, L.** Phys. Rev. **37:** 405. 1931; **38:** 2265. 1931.
[3] **Thomson, W.** (Lord Kelvin). Proc. Roy. Soc. Edinburgh **1854:** 123
[4] **Helmholtz, H. V.** Wied. Ann. **3:** 201. 1878.
[5] **MacInnes, D. A., & J. A. Beattie.** Jour. Am. Chem. Soc. **42:** 1117. 1920.

The latter is a homogeneous quadratic function of the rates of flow

$$2F(\mathbf{J}, \mathbf{J}) \equiv \sum_{i,k} R_{ik}(\mathbf{J}_i \cdot \mathbf{J}_k) \tag{7}$$

so constructed that it always equals the rate of conversion of free energy into thermal energy (like Joule heat, for example). We naturally adopt the convention

$$R_{ik} = R_{ki} \tag{7a}$$

By the second law of thermodynamics we have the inequality

$$2F \geqq 0$$

The = sign applies to the special case that all velocities are equal:

$$F((c_i\mathbf{v}), (c_i\mathbf{v})) = 0 \tag{8}$$

for moving a container with solution in it causes no dissipation of energy. In terms of the coefficients R_{ik} this condition becomes

$$\sum_k R_{ik}c_k = 0 \tag{7b}$$

It may be presumed, however, that no *relative* motion of the components in a solution can take place without dissipation of energy, so that the function (7) does not normally vanish except in the case (8).

We shall use the following notations for the thermodynamic quantities:

$$Z = U - TS + PV \tag{9}$$

for the energy at constant pressure and

$$\mu_i = (\partial Z/\partial N_i)_{P,T} \tag{10}$$

for the thermodynamic potentials.

Under the conditions of constant pressure and (substantially) constant temperature we have

$$-dZ/dt = -\int \sum_i \mu_i(\partial c_i/dt)dV$$
$$= \int \sum_i \mu_i(\nabla \cdot \mathbf{J}_i)dV = -\int \sum_i (\mathbf{J}_i \cdot \nabla\mu_i)dV \tag{11}$$

assuming that the outward components of the vectors $\mathbf{J}_1, \mathbf{J}_2, ..$ vanish at the boundaries of the liquid system.

Now the laws of diffusion may be summarized in terms of a variation-principle: *The fields of flow* $\mathbf{J}_1(x,y,z)$, $\mathbf{J}_2(x,y,z)$, .. *will arrange*

themselves in such a manner that

$$- (dZ/dt) - \int F(\mathbf{J}, \mathbf{J})\, dV = \text{Maximum} \tag{12}$$

With the aid of Equation (11) we get

$$\sum_i (\delta \mathbf{J}_i \cdot \nabla \mu_i) + \delta F(\mathbf{J}, \mathbf{J}) = 0 \tag{13}$$

to be satisfied locally, and, substituting Equation (7)

$$- \nabla \mu_i = \sum_{k=1}^{s} R_{ik} \mathbf{J}_k \tag{14}$$

On account of Equations (7a) and (7b), the number of independent coefficients is only $\frac{1}{2} s(s-1)$, as compared to the $(s-1)^2$ coefficients of Equation (3) with corollaries (3a, 3b). The system (14) thus contains $\frac{1}{2}(s-1)(s-2)$ reciprocal relations which are not trivial consequences of the set (7b).

The connection between Equations (14) and (3) can be obtained by straight substitution of the latter in (14). This yields

$$\sum_{j,k} R_{ij} D_{jk} \nabla c_k \equiv \nabla \mu_i + K A_i \nabla P \equiv \sum_k \left(\frac{\partial \mu_i}{\partial c_k} + A_i \bar{V}_k \right) \nabla c_k \tag{15}$$

The extra term involving ∇P must be admitted because the identities need only be true at constant pressure. The actual values of $A_1, \ldots A_s$ are determined by the conventional relation (3b), as follows

$$0 = \sum_{j,k} R_{ij} D_{jk} c_k = \sum_k \left(\frac{\partial \mu_i}{\partial c_k} + A_i \bar{V}_k \right) c_k$$

$$= \sum_k \left(\frac{\partial \mu_k}{\partial c_i} + A_i \bar{V}_k \right) c_k = \frac{\partial P}{\partial c_i} + A_i = (\bar{V}_i / K) + A_i$$

where K denotes the compressibility as before. Thus, thanks to the convention (3b) we may write Equation (15) in the form

$$\sum R_{ij} D_{jk} = \frac{\partial \mu_i}{\partial c_k} - \frac{\bar{V}_i \bar{V}_k}{K} = V \left(\frac{\partial \mu_i}{\partial N_k} \right)_P \tag{16}$$

For a two-component system described by Equation (1) this reduces to

$$D R_{ik} = V (\partial \mu_i / \partial N_k)_P \tag{16a}$$

The conditions which are imposed by the reciprocal relations (7a) upon the coefficients of Equation (3) are obtained most easily from the observation that the sum

$$\sum_{j,l} R_{jl} D_{ji} D_{lk}$$

is a symmetrical function of the fixed indices (i, k). Observing Equation (3a) we then obtain from Equations (16)

$$\sum_j \left(\frac{\partial \mu_i}{\partial N_j}\right)_P D_{jk} = \sum_j \left(\frac{\partial \mu_k}{\partial N_j}\right)_P D_{ji} \tag{17a}$$

or

$$\sum_j \frac{\partial \mu_i}{\partial c_j} D_{jk} = \sum_j \frac{\partial \mu_k}{\partial c_j} D_{ji} \tag{17b}$$

These relations, then, describe the connection between the diffusion in a system of 3 or more components and the thermodynamic properties of the solutions. When the latter are known, the relations (17) serve to diminish the number of independent coefficients D_{ik}. Conversely, one may derive some information of thermodynamic nature from studies of diffusion, although present experimental technique is not so good that this procedure would lend itself to accurate measurements of thermodynamic functions.

Even for systems of unknown colligative properties, we can draw certain qualitative conclusions of some interest.

Let us consider a case of differential diffusion, that is, the initial inequalities of composition are so small that the consequent variations of the coefficients D_{ik} may be neglected. Then the diffusion is governed by a set of partial differential equations with constant coefficients, namely the system (3) together with the equations of continuity

$$\frac{\partial c_i}{\partial t} + (\nabla \cdot \mathbf{J}_i) = 0 \tag{18}$$

supplemented by the boundary conditions:

$$(\mathbf{n} \cdot \mathbf{J}_i) = 0 \tag{19}$$

(where $\mathbf{n}$ denotes the normal of the boundary).

Usually such a system of differential equations may be separated by a linear transformation of the variables

$$c_i - \bar{c}_i = \sum_k A_{ik} y_k \tag{20}$$

to yield individual equations of the form

$$\frac{\partial y_i}{\partial t} = B_i \nabla^2 y_i \tag{21}$$

although in general certain complications may occur. However, the

substitution (20) may be characterized by the requirement that it transform the dissipation-function (7) and the Hessian of the free energy:

$$V \sum_{i,k} \left(\frac{\partial^2 \psi}{\partial N_i \partial N_k}\right)_V \Delta c_i \Delta c_k = \sum_{i,k} \left(\frac{\partial \mu_i}{\partial c_k}\right) \Delta c_i \Delta c_k \tag{22}$$

simultaneously to sums of squares. Now the form (22) is definite and the form (7) is semi-definite. It is known that under these conditions we can always find a transformation (20) which leads to a set of simple equations of the type (21). Moreover, the coefficients A_{ik} and B_i are real and positive. The latter result implies that *the relaxation-periods for diffusion are always real;* this is but a special case of a rule which applies generally to any process whose course is determined by a dissipation-function.

Looking back at our analysis, we see that, in the general case, the way to describe the properties of a solution as regards diffusion most simply and compactly is to specify the dissipation-function (7). For a system of only two components, it is admittedly more practical to specify the single coefficient of Fick's law; but the advantages of that scheme cannot be preserved in any generalization to cases which involve a greater number of components This observation suggests that the significance of such correlations as may exist between diffusion and other phenomena will appear most clearly when they are described in terms of the dissipation-function.

The coefficient of diffusion in Equation (1) is thus considered to be a product of two factors, the other one being given by the variation of the thermodynamic potentials with the composition. This has a direct significance in terms of kinetic theory.

According to our kinetic picture of diffusion, the molecules exchange places unceasingly whether or not a concentration gradient is present The frequency of this exchange is directly related to the dissipation-function[2]; the coefficients of the form (7) are inversely proportional to that frequency.

A systematic relative motion of two kinds of molecules in solution results from the random exchange only if a concentration gradient exists in the first place, so as to create a statistical bias. The effect of a given gradient may be modified by intermolecular forces which cause deviations from the laws of ideal solutions. These effects are properly accounted for by Boltzmann's relation between entropy and probability. As a result, the gradients of the thermodynamic potentials deter-

mine the ratios of the numbers of molecules being displaced in opposite directions.

Diffusion and Hydrodynamics

Viscous flow is a relative motion of adjacent portions of a liquid. Diffusion is a relative motion of its different constituents.

Strictly speaking, the two are inseparable; for the "hydrodynamic" velocity in a diffusing mixture is merely an average determined by some arbitrary convention. For certain practical reasons, we have adopted the definition (4); but one should not infer that precisely the gradient of this "average" velocity determines the viscous transfer of force.

While these questions have never been analyzed thoroughly, for ordinary purposes they are probably of small practical importance, as is the analogous question about the "compression viscosity" of a compressible fluid. As an example, we may consider a solution streaming through a narrow capillary. A simple analysis shows that the differential motion of the various constituents of the solution will be proportional to the pressure gradient. The ratio of the differential velocity to that of the main flow is then of the order (a^2/r^2), where r is the radius of the capillary and a some fixed length, presumably of molecular dimension, determined by the composition of the solution and the nature of the capillary wall. The reciprocal effect is well known. It is just the osmotic pressure difference maintained across a more or less imperfect semipermeable membrane. As sources of error in the study of diffusion by means of porous discs, these "filter" effects are not as important as the modifications of composition near the capillary walls, which are obviously of the order (a/r). As to whether the coupling of viscous flow and diffusion might play a role in physiological processes, nothing is known.

DIFFUSION OF NON-ELECTROLYTES

Dilute Solutions

The kinetic problem of diffusion is simplified in some respects by specialization to dilute solutions. We are then dealing with a population of solute molecules—of varying density but everywhere sparse—in a practically constant environment of solvent. In some respects, the molecular structure of the solvent does not matter. The relevant thermodynamic properties of the solution are described by the well-known rule for ideal solutions

$$\mu_i - \text{const.} = kT \log (N_i/N_1); \qquad (i = 2, 3, \cdots) \tag{23}$$

whence

$$V(\partial\mu_i/\partial N_i)_{P,T} = kTV/N_i = kT/c_i; \qquad (i = 2, 3, \cdots)$$
$$V(\partial\mu_i/\partial N_j)_{P,T} \sim 0; \qquad (j \neq i) \tag{24}$$

Moreover, we may presume that the various solute molecules will have constant individual coefficients of diffusion D_2, D_3, ..., relative to the solvent; and, for most purposes, it is not necessary to distinguish between the hydrodynamic "velocity of the solution" and the velocity of the solvent. The dissipation-function will be of the form

$$2F = \sum_{i=2}^{s} R_{ii} |\mathbf{J}_i - c_i\mathbf{v}_1|^2 = \sum_{i=2}^{s} c_i^2 R_{ii} |\mathbf{v}_i - \mathbf{v}_1|^2 \tag{25}$$

Its coefficients are related to the coefficients of diffusion and to the molecular "coefficients" of friction" ρ_2, ρ_3, ... as follows

$$c_i R_{ii} = \rho_i = kT/D_i, \tag{26}$$

and ρ_2, ρ_3, ... are independent of the concentrations (approach finite limits for low total concentrations of solutes).

For particles which are large compared to the solvent molecules, the coefficients of friction can be computed from the laws of hydrodynamics, e.g. for spheres[6] of radius a

$$\rho = 6\pi\eta a \tag{27}$$

(where η denotes the viscosity of the solvent) and for ellipsoids of revolution[7] with semi-major axes a, a, b:

$$\rho = 6\pi\eta(b^2 - a^2)^{1/2}/\cosh^{-1}(b/a) \tag{28}$$

Equation (28) is in accord with the expectation that the spherical shape will yield the least resistance (as a harmonic average over orientations) for a given molecular volume V_M. If this conjecture is correct, we shall have the general inequality

$$\rho \geqq 6\pi\eta(3/4\pi)^{1/3}(V_M/N_A)^{1/3} \tag{29}$$

valid to the extent that the hydrodynamic picture is reliable. For shapes which do not differ greatly from the spherical, Equation (29) may be read as an approximate equality.[7]

The given hydrodynamic relations are useful when we have to interpret the diffusion of colloids or large molecules such as proteins, even sugars (in aqueous solution).

---

[6] Stokes, G. Cambridge Phil. Soc. Trans. 9: 5. 1856.
[7] Perrin, F. Jour. de Physique et le Radium 7: 1. 1936.

In cases where the molecules of solute and solvent are of comparable size, we may expect to find relations which differ from those given by the hydrodynamic theory, presumably more the smaller the relative dimensions and masses of the solute molecules. On the basis of available data, the coefficients of diffusion are indeed consistently greater than the maximum allowed by the inequality (29). In the tables compiled by Kincaid, Eyring and Stearn[8] the deviations vary from 25 per cent to a maximum of 250 per cent.

The extreme figure refers to the diffusion of bromoform in amyl alcohol. High values are also found for that solute in other alcohols. The molecular dimensions of these solvents are not very large, but it is a tenable point of view that, for the purpose in hand, we should assume that the alcohols are effectively polymerized.

Where the interdiffusion of reasonably non-polar molecules is concerned, one may—with a bit of good will—read from the data a measure of correlation between the function

$$D\eta V_M^{1/3}/T$$

and the ratio

$$V_{\text{solvent}}/V_{\text{solute}}$$

but there is no fixed correspondence. Indeed, it is difficult to formulate any generalization from the data beyond the simple statement that the ratio

$$l = kT/D\eta \tag{31}$$

is a length of the order of magnitude of molecular dimensions, normally smaller than the value $6\pi a$ given by Equation (27).

From the point of view of molecular theory, viscous flow and diffusion present parallel problems. It would seem that for an exact theory of either, we should have to analyze the cooperative character of the molecular motion involved; but this difficult analysis has not yet been developed further than the hydrodynamic approximation.

In Eyring's approach to the kinetic theory of liquids,[9, 10] the smallest possible number of degrees of freedom are considered together. Thus, the strictly kinetic aspects of the problems involved are treated on a level comparable to that of elementary gas theory; but the approach is much more sophisticated in that the statistical interpretation of the

[8] Kincaid, J. F., H. Eyring, & A. E. Stearn. Chem. Rev. 28: 301. 1941.
[9] Eyring, H. Jour. Chem. Phys. 4: 283. 1936.
[10] Roseveare, W. E., R. E. Powell, & H. Eyring. Jour. Applied Phys. 12: 669. 1941.

thermodynamic functions is fully exploited. The method lends itself to a facile interchange of deduction and induction. The theorist's haughty ideal of doing without all inductive elements has not yet been obtained; but, even so, we can say that, on the whole, we now understand the laws which determine the rates of viscous flow and diffusion, at least in dilute solutions of non-polar liquids.

We shall review some of the results which have been brought out largely with the aid of Eyring's theory and point out some outstanding problems. For comfort in traveling, we shall try to retain an inductive point of view.

A very significant general rule was discovered by Bridgman[11] who investigated the viscosities of liquids at high pressures. He found that, for non-polar liquids, the temperature coefficients of viscosity *at constant volume* are small, so that the variation of viscosity with temperature at constant pressure is mostly due to the thermal expansion. This rule does not apply to polar liquids; water is a striking example. Bridgman's rule means that the successive changes of configuration of the molecules which are involved in over-all deformations (viscous flow) can and do take place without passing through configurations whose potential energy exceed the normal by more than about $2kT$. This statement refers to the condition of constant total volume, whereby the change of potential energy due to a local expansion will be compensated by the consequent compression in other parts of the liquid Bridgman's rule invites the interpretation that the energy of activation for each single step in the process of viscous flow measures the temporary local expansion required for that step. The relative expansions involved turn out to be comparable for different liquids. In order to predict absolute viscosities, one has to deal with certain subtle geometrical questions such as the extent of deformation accomplished by each local rearrangement of the molecules. A simple picture suggested by Hirschfelder, Stevenson and Eyring[12] leads to results of the observed order of magnitude.

An elementary step in diffusion is pictured simply as a jump of a single molecule from one equilibrium position to another. The predicted relation between viscosity and diffusion comes out in the form

$$kT/D\eta = \lambda_2\lambda_3/\lambda_1 \tag{32}$$

where λ_1, λ_2, λ_3, are lengths of molecular dimensions.

[11] **Bridgman, P. W.** The Physics of High Pressures. The Macmillan Company, New York. 1931.

[12] **Hirschfelder, J. O., D. P. Stevenson, & H. Eyring.** Jour. Chem. Phys. 5: 896. 1937.

The experimental results can be fitted by adjustments of the parameters λ between reasonable bounds. It does not seem possible to obtain more accurate predictions without considerable refinement of the theory.

An interesting qualitative rule was formulated on empirical grounds by Öholm[14]: The solutes which diffuse most rapidly have the smallest temperature coefficients for diffusion. The rule can be interpreted very simply by the plausible assumption (inherent in Eyring's general theory) that the molecule which needs less energy to move from one position of relative equilibrium to another will do so more often.

Concentrated Solutions

In dealing with concentrated solutions, we need to know the colligative properties of the system in hand as well as the coefficient of diffusion before we can compute the energy dissipated by the process. In this sense, only a few systems have been investigated properly.

Kincaid, Eyring and Stearn[8] analyzed data from several systems and ventured some generalizations:

(1) For systems which obey Raoult's law (benzene-chloroform), the coefficient of diffusion varies rather simply with the concentration in that the product $(D\eta)$ is a linear function of the composition

$$D\eta/kT = \beta_{21}x_1 + \beta_{12}x_2;$$

(x_1, x_2 = mol-fractions).

(2) For some solutions of non-polar liquids which exhibit considerable deviations from Raoult's law, the quotient

$$D\eta/(x_1\partial\mu_1/\partial x_1) = \beta_{21}x_1 + \beta_{12}x_2 \tag{33}$$

(μ = chemical potential) varies linearly with the concentration. The former relation is but a special case of this.

(3) For some mixtures of polar liquids, such as the pair water-ethyl alcohol, the left member of Equation (33) assumes values much larger than those derived from linear interpolation between the two extremes of dilute solutions.

In the cases studied by the authors, the ratios (β_{12}/β_{21}) did not differ much from unity, so that the exact type of interpolation selected does not set an important precedent.

[14] Öholm, L. W. Medd. K. Vetenskapsakad. Nobelinstitut 2: 16. 1912.

Equation (33) is in effect a proportional relation between the viscosity and the coefficients of the dissipation-function for diffusion, as follows:

$$c_2/c_1R_{11} = -1/R_{12} = c_1 \; c_2R_{22}$$
$$= (\beta_{21}x_1 + \beta_{12}x_2)(c_1 + c_2)/\eta \qquad (34)$$

The relation, which has been derived from a very limited number of examples, ought to be tested more extensively.

As regards the more complicated cases of type 3, solutions of alcohol and water are much more viscous than either component alone. As pointed out by Kincaid, Eyring and Stearn, the rule expressed by Equation (33) fails here in the sense that an effect which impedes viscous flow does not impede the diffusion, or not as much.

The different behavior of such systems as alcohol-water may be due to the polar nature of the components. This is quite plausible although no detailed mechanism has been suggested. However, the evidence on hand is far from conclusive, for the only examples given of systems which obey Equation (33) show either negative deviations from Raoult's law, or none.

Negative deviations from Raoult's law correspond to alternating arrangements of the molecules, which would tend to increase the dissipation-function for diffusion. *Any* systematic structure is apt to interfere with viscous flow, which accounts qualitatively for the observed correlation.

On the other hand, a positive deviation from Raoult's law corresponds to local segregation of the components, which might well facilitate diffusion and yet impede viscous flow.

Accordingly, a study of diffusion in systems of non-polar liquids which exhibit positive deviations from Raoult's law would help to fill an important void in our present knowledge.

In this connection, a most interesting question is what may happen to the coefficient of diffusion in a liquid system with a critical point Presumably, it will vanish because the gradient of the thermodynamic potential does. However, the fluctuations of composition become very large, and it seems quite conceivable that the coefficients of the dissipation-function may vanish too, although not necessarily of the same order.

SOLUTIONS OF ELECTROLYTES

In one way, electrolytes are most obliging solutes: To a first approximation, their properties are additive functions, not merely of the

components, but of the very ions. However, their deviations from the ideal additive behavior decrease much more slowly with decreasing concentration than those of other solutes. Fortunately, this peculiarity can be accounted for in terms of electrostatic forces, which may be trusted to obey Coulomb's law for distances that are large on a molecular scale. The persistent parts of the variations, which are generally proportional to $C^{1/2}$, are due to the long reach of the Coulomb forces This means that, whenever an effect is proportional to $C^{1/2}$ in the limit, its coefficient can be predicted from the charges and limiting mobilities of the ions present. The colligative properties, the migration of the ions in electric fields, the coefficients of diffusion and the viscosities of electrolytic solutions all exhibit the linear variation with $C^{1/2}$. These effects have all been computed theoretically[14]; the predictions have been generally verified by experiment with the one exception that the evidence as concerns diffusion is incomplete.

Improved approximations often bring in terms which vary as $c \log c$. The coefficients of these are also predictable; the computations have been carried out for the colligative properties of "unsymmetrical" electrolytes[15] and for the surface tensions of binary electrolytes.[14] The corresponding computation for diffusion has been dealt with in part.

Beyond these limiting laws, the properties of electrolytes, even of the same valence type, etc., exhibit differences which are proportional to the first and higher powers of the concentration These specific differences depend on the short range interaction of the ions. The electrostatic forces are important but the over-all result cannot be predicted on the basis of Coulomb's law alone. Solutions of neutral molecules exhibit analogous individual differences; but among electrolytes the variety is much greater because the forces involved are generally stronger.

To the first approximation of additive properties, the dissipation-function for diffusion and the pertinent thermodynamic properties are given by Equations (24) and (25), respectively, as for neutral molecules. There is this difference that Equation (24) involves one redundant variable; for the concentrations of individual ions are subject to the condition of electrical neutrality:

$$\sum c_i e_i = 0 \tag{35}$$

[14] An excellent systematic presentation, complete except for a few specialized topics of minor practical importance, is given by **Harned, H. S., & B. B. Owen.** "The Physical Chemistry of Electrolytic Solutions." A C.S. Monograph No. **95.** Reinhold Publishing Corporation, New York, N. Y. 1943

[15] **LaMer, V. K., & [illegible] Mason.** Jour. Am. Chem. Soc. **49:** 420 1927.

Moreover, the solution may be subjected to an electric field. Equation (25) actually represents the dissipation-function for conduction and diffusion combined. Its coefficients can be derived from measurements of electrical conductivity together with the determination of one electrolytic transference number. The remarkable fact that the coefficients of diffusion of electrolytes can be predicted thus from their electrical properties was first recognized by Nernst [16] From this indirect source we know much more about the limiting coefficients of diffusion of ions than we know about the diffusion of neutral molecules.

The coefficients of diffusion of ions are normally smaller than those of molecules of comparable dimensions. The electrolytic mobilities of most ions are about equal to the values computed from Stokes' formula, sometimes a little greater,[17] often much smaller. In water, the small elementary ions of Li, Na, F migrate more slowly than those of K, Rb, Cs, Cl, Br and I. For Li^+, the slowest of all, the variation with the temperature corresponds to that of the fluidity. The others have smaller temperature coefficients in inverse order of their speeds. The coefficients of diffusion of more highly charged ions are equal to that of Li^+ or smaller. These general facts are to be interpreted somehow in terms of the electrostatic interaction between solute and solvent. The forces involved are greater than ordinary intermolecular forces, and more so the smaller the size of the ion. On the other hand, a very large ion can only move as fast as the solvent molecules can get out of the way; thus, we can understand why there should be an optimum ion size for greatest speed. In solvents other than water, e.g., alcohols, larger ions are often the fastest. In acetone,[18] the equivalent conductivities of Li, Na and K ions at 25° C. are all about equal to 70; that of the tetramethylammonium ion is 102.8.

An extension of Eyring's theory to deal with these phenomena might be profitable; but such a development must perhaps wait for some further advances in the theory of polar liquids. Some questions connected with the energies and entropies of solvation of ions are also fundamental to the kinetics of diffusion. For an inductive approach, a comprehensive analysis of kinetic and colligative properties together may be the best plan.

[16] **Nernst, W.** Zeit. physik. Chem. **2**: 613. 1888.

[17] The exceptionally high mobilities of hydrogen and hydroxyl ions in water and of the former in some other solvents must be ascribed to a special proton jump mechanism. The theoretical problem is difficult. See **Hückel, E.** Zeit. Elektrochem. **34**: 546. 1928; **Bernal, J. D., & R. H. Fowler.** Jour. Chem. Phys. **1**: 515. 1933; **Wannier, G.** Ann. d. Physik **24**: 545. 1935.

[18] **Walden, P., H. Ulich, & G. Busch.** Zeit. Physik. Chem **123**: 429. 1926.

Diffusion and Electrolytic Conduction

Since we recognize the ions as kinetic units, it is logical to describe diffusion and electrolytic conduction together in one scheme. In fact, any attempt to describe diffusion separately would involve artificial restrictions, for diffusion in a phase which contains at least one electrolyte among three or more components will cause electric currents unless special conditions are fulfilled.

In several ways, the simplest scheme is to construct a dissipation-function for the relative motion of all ions and molecules present, of the general form given by Equation (7). The variation principle (12) then applies with the modification that the work of applied electric potential differences $(\varphi_p - \varphi_q)$ must be included:

$$\sum_q i_q\varphi_q - \frac{dZ}{dt} - F(\mathbf{J}, \mathbf{J}) = \text{Maximum}, \tag{12a}$$

where i denotes the current entering at the q'th electrode, and dZ/dt is meant to include free energy changes at the electrodes. Equation (35) (differentiated with respect to the time) must be taken into account as a restriction. Its Lagrangian multiplier $\varphi(x, y, z)$ represents the electrostatic potential. Equations (14) then result with the convention that for all ions μ_i shall mean the *total potentials*

$$\mu_i = (\mu_i)_{\text{chemical}} + e_i\varphi$$

also referred to as the "electrochemical potentials." We thus avail ourselves of a device invented by Brönsted[19] and Guggenheim[20] in order to circumvent the totally elusive problem of defining separately the electrostatic potential and the chemical potentials of the individual ions

In addition to the reciprocal relations which apply to diffusion alone, this scheme of description assumes and implies the commonly accepted analogous relations between electrolytic transference numbers and the electromotive forces of cells with liquid junctions.[4, 5]

The Effects of Coulomb Forces

When a charged body is immersed in a conductor, the mobile charges in the conductor will rearrange themselves so as to screen off the field due to the immersed charge (to the extent that the charges are not neutralized by combination). When the conductor is a solution of ions, the screening follows an exponential law and the mean distance

---

[19] Brönsted, J. N. Zeit. Physik. Chem. 143: 301. 1929.
[20] Guggenheim, E. A. Jour. Phys. Chem. 33: 842. 1929; 34: 1540 1930.

$1/\kappa$ of the screening charge can be computed from statistical considerations which lead to the general formula

$$\kappa^2 = \frac{4\pi}{\epsilon kT} \sum_i c_i e_i^2 \tag{36}$$

where ε denotes the dielectric constant and $e_1, e_2, \ldots$ the charges of ions present in the concentrations $c_1, c_2, \ldots$ (ions/cm³).

Debye and Hückel[21] recognized that the electric fields of the ions themselves are similarly screened. For the potential of the average field surrounding an ion of charge e_j, they derived the formula

$$\psi_j(r) = (e_j/\epsilon)e^{-\kappa r}/r \tag{37}$$

For the local concentrations of other ions at a distance r they found

$$c_{ji}(r) = c_i e^{-e_i\psi_j/kT} \sim c_i\left(1 - \frac{e_j e_i}{kT}\frac{e^{-\kappa r}}{r}\right) \tag{38}$$

which leads to the average electric charge density

$$\rho_j(r) = -e_j(\kappa^2/4\pi)e^{-\kappa r}/r \tag{39}$$

According to Equation (37) the mean potential at an ion due to its "atmosphere" of compensating charge equals

$$\psi_j^*(0) = -\kappa e_j/\epsilon \tag{40}$$

This leads to the limiting law for the thermodynamic potentials

$$\mu_j = \text{const.} + kT \log c_j - (\kappa e_j^2/2\epsilon) \tag{41}$$

Various semi-empirical formulas which reduce to Equation (41) in the limit have been found to describe the thermodynamic properties accurately over a considerable range of concentrations, for example[14]

$$\mu_j = \text{const.} + kT \log c_j - \frac{\kappa e_j^2}{2\epsilon(1+\kappa a)} + Bc \tag{42}$$

which can be applied to many univalent binary electrolytes in aqueous solution; here a and B are adjustable constants whose physical meanings are understood to a degree.

Debye and Hückel showed that the Coulomb forces between the ions will affect their migration velocities by two concurrent mechanisms.[22] Their computations were subsequently improved and generalized by other authors.[23, 24]

One effect arises from the volume force which the external applied

[21] Debye, P., & E. Hückel. Physik. Zeit. 24: 185. 1923
[22] Debye, P., & E. Hückel. Physik. Zeit. 24: 305. 1923
[23] Onsager, L. Physik. Zeit. 27: 388. 1926; 28: 277. 1927
[24] Onsager, L., & R. M. Fuoss. Jour. Phys. Chem. 36: 2689. 1932

field (electrostatic or other) exerts upon the ions in the atmosphere. This force is transferred to the solvent and causes hydrodynamic flow. In the case of electrical conduction, the density of the volume force is proportional to the local charge density (Equation 39). As a result each ion has to move against a local counter-current of velocity

$$\Delta v_j = -\kappa e_j X / 6\pi\eta \tag{43}$$

where X denotes the intensity of the electric field. For diffusion we need a more general result, derived[24] for the case that a set of force fields of intensities $\mathbf{k}_1$, $\mathbf{k}_2$, ... are acting upon the respective kinds of ions, with a compensating field of force $\mathbf{k}_0$ acting upon the solvent molecules. In the ionic atmosphere, the equilibrium of this system of forces is disturbed because the concentrations differ from the average concentrations in the solution. The resultant volume force causes a motion of the liquid which imparts to every ion of charge e_j the additional velocity

$$\Delta \mathbf{v}_j = -(2e_j/3\eta\kappa\epsilon kT) \sum_i c_i e_i \mathbf{k}_i \tag{44}$$

which reduces to Equation (43) when the forces $\mathbf{k}_i$ are proportional to the charges e_i. Since κ is proportional to the square root of the concentration, so is the effect described by Equation (44): $c^{-\frac{1}{2}}\ c = c^{\frac{1}{2}}$. The effect vanishes—simultaneously for all ions—whenever the condition

$$\sum_i c_i e_i \mathbf{k}_i = 0 \tag{45}$$

happens to be satisfied, as in the diffusion of an electrolyte whose ions have the same individual coefficients of diffusion.

To improve the approximation expressed by Equation (44), Onsager and Fuoss[24] retained one more term in the power series expansion of Equation (38), thus

$$c_{ji}(r) \sim c_i \left(1 - \frac{e_j e_i}{\epsilon kT} \frac{e^{-\kappa r}}{r} + \left(\frac{e_j e_i}{\epsilon kT}\right)^2 \frac{e^{-2\kappa r}}{2r^2}\right)$$

and computed a correction term which varies as $c \log c$ in the limit of low concentrations. This term always increases the coefficient of diffusion.

The second effect recognized by Debye and Hückel[22] arises when the centrally symmetrical arrangement of the charges in the ionic atmospheres is disturbed by migration of the ions. It is true that the interplay of Coulomb forces and thermal motion tends to restore the symmetrical equilibrium distribution, but this process needs a finite

time. The *relaxation time* for the readjustment of electric charges in a conductor is given by Maxwell's formula

$$\tau = \epsilon/4\pi\lambda \sim \rho/\kappa^2 kT \tag{46}$$

(λ = conductivity).

Alternatively, we may estimate the time needed for diffusion to equalize concentration in a region of linear dimension s

$$\tau \sim s^2/D = \rho s^2/kT \tag{47}$$

For $s \sim 1/\kappa$ the two estimates are identical as to order of magnitude.

Now the force due to the disturbance of the atmosphere by migration can be estimated as follows: In the time τ, the ion moves a distance $\mathbf{v}\tau$, which results in a relative distortion of the order $\mathbf{v}\tau/(1/\kappa) = \kappa\mathbf{v}\tau$. The "total" force between ion and atmosphere is of the order $\kappa^2 e^2/\epsilon$. (The region nearest to the ion contributes more but by Equation (47) the relaxation time for this region is shorter.) On this basis we arrive at the estimate

$$-\Delta\mathbf{k} \sim \mathbf{v}\kappa e^2/\epsilon kT \tag{48}$$

for the "relaxation effect." Detailed computation verifies the estimate apart from a variable numerical factor, but with the modification that the force also depends on the mean motions of the other ions.

The relaxation effect can also vanish. This happens for all the ions when they all have the same average motion:

$$\begin{gathered} \Delta\mathbf{k}_1 = \Delta\mathbf{k}_2 = \cdots = \Delta\mathbf{k}_s = 0; \\ (\mathbf{v}_1 = \mathbf{v}_2 = \cdots = \mathbf{v}_s) \end{gathered} \tag{49}$$

This result is almost obvious without computation. When all the ions move with the same speed, the whole systematic arrangement of atmospheres moves along undisturbed and no dissymmetry arises.

We note that, in the diffusion of a simple electrolyte (two ions) with no electric current flowing, the conditional clause of Equation (49) is automatically satisfied, and the relaxation forces vanish.

It has been shown[25, 26, 24] that a similar relaxation effect also modifies the viscosities of electrolytes. The order of magnitude of this effect can be estimated as follows: The rate of deformation is given by the velocity gradient

$$\nabla\mathbf{v};$$

---

[25] **Jones, G., & M. Dole.** Jour. Am. Chem. Soc. **51:** 2950. 1929.
[26] **Falkenhagen, H., & M. Dole.** Zeit. physik. Chem. **6:** 159. 1929; Physik. Zeit. **30:** 611 1929. **Falkenhagen, H.** Physik. Zeit. **32:** 365. 1931. **32:** 745. 1931.

the unrelaxed deformation

$$\tau \nabla \mathbf{v}$$

gives us the directed fraction of the total transport of force

$$(e^2\kappa^2/\epsilon)\kappa^{-1} = e^2\kappa/\epsilon$$

in the ionic atmosphere. We thus estimate an additional viscosity of the order

$$\Delta\eta \sim \tau c e^2 \kappa/\epsilon \sim \kappa\rho \tag{50}$$

Detailed computation yields for the case that all the ρ_i are equal

$$\Delta\eta = \kappa\rho/480\,\pi \tag{51}$$

The effect is proportional to $c^{1/2}$, with a fairly small numerical factor. In fact, the relative increase of the viscosity is just 1/80 of the relative increase of the electrical resistance due to the electrophoretic effect alone; cf. Equation (43).

It is, nevertheless, interesting to inquire—as a matter of principle—how the "electrostatic" viscosity will, in turn, affect the coefficients of diffusion of the ions.

A simple consideration will show that this secondary effect is of a smaller order of magnitude. The main point is that the estimate given by equation (50) is only good for viscous flow which deforms relatively large regions, at least of the order $1/\kappa$. For smaller regions, we must use the short relaxation time which we get from Equation (47) when $s < 1/\kappa$. Now the frictional resistance to the motion of an ion is mainly due to the deformation of the liquid in its immediate vicinity. The deformation of the region $r > 1/\kappa$ contributes only an amount of the order κ/η to the mobility. On this basis, we estimate that for the motion of ions (or any small molecules), the electrostatic viscosity is but a very small relaxation effect of a higher order, proportional to the first power of the concentration (κ^2).

This little consideration illustrates the dubious nature of the "correction" which we apply to diffusion and mobility data on account of the viscosity. It is true that this procedure often clarifies other correlations, but it is only a makeshift which has to serve because we do not know any better. The part of the viscosity which arises from the electric relaxation effect should preferably be eliminated before the correction is applied. Of course, it makes very little difference, except possibly for the purpose of extrapolation.

Before we summarize the results, we may as well admit that the theory is not accurate enough to justify a distinction between the veloci-

ties of the ions relative to the "solution at rest" and their velocities relative to the solvent. The theory is, at best, on the level of accuracy of Equation (25) for neutral molecules. For one who desires formulas which satisfy Equation (7b), it will be permissible in the following to replace $\mathbf{J}_i$ throughout by $(\mathbf{J}_i - c_i\mathbf{v}_0)$ where $\mathbf{v}_0$ denotes the velocity of the solvent.

We first assemble the dissipation-function for combined conduction and diffusion of a simple electrolyte

$$\begin{aligned}
2F &= R_{11}\mathbf{J}_1^2 + 2R_{12}(\mathbf{J}_1 \cdot \mathbf{J}_2) + R_{22}\mathbf{J}_2^2 \\
&= (\rho_1/c_1)\mathbf{J}_1^2 + (\rho_2/c_2)\mathbf{J}_2^2 \\
&\quad + \frac{-e_1e_2\rho_1\rho_2\kappa}{3\epsilon kT}\,\frac{(1 - q^{1/2})}{(c_1e_1^2\rho_1 + c_2e_2^2\rho_2)}\,|e_1\mathbf{J}_1 + e_2\mathbf{J}_2|^2 \\
&\quad + \frac{2}{3\eta kT\epsilon(1 + \kappa a)\kappa}\,|e_1\rho_1\mathbf{J}_1 + c_2\rho_2\mathbf{J}_2|^2 \\
&\quad - \frac{\varphi(\kappa a)}{3\eta(\epsilon kT)^2}\,|e_1^2\rho_1\mathbf{J}_1 + e_2^2\rho_2\mathbf{J}_2|^2
\end{aligned} \tag{52}$$

where a stands for the least distance of approach of the ions and

$$q = (c_1e_1^2\rho_2 + c_2e_2^2\rho_1)/(c_1e_1^2 + c_2e_2^2)(\rho_1 + \rho_2)$$

$$\varphi(\kappa a) = e^{2\kappa a}(1 + \kappa a)^{-2}\int_{2\kappa a}^{\infty} e^{-t}\,dt/t \tag{52a}$$

The first two terms represent the dissipation-function without allowance for the effects of the Coulomb forces. The third term is due to the relaxation effect. It vanishes if there is no electric current, for the reason explained above. The last two terms represent the electrophoretic effects of first and second order, with approximate allowances for the finite dimensions of the ions.

To specialize Equation (52) for diffusion, we put

$$\mathbf{J}_1/c_1 = \mathbf{J}_2/c_2 = \mathbf{v}$$

which yields

$$2F = \left(c_1\rho_1 + c_2\rho_2 + \frac{2}{3\eta\epsilon kT}\,\frac{(c_1e_1)^2}{\kappa(1 + \kappa a)}\,(\rho_1 - \rho_2)^2 \right.$$
$$\left. \frac{\varphi(\kappa a)(c_1e_1)^2}{3\eta(\epsilon kT)^2}\,(e_1\rho_1 - e_2\rho_2)^2\right)\mathbf{v}^2 \tag{53}$$

The relaxation effect vanishes; the electrophoretic effects remain. These may be interpreted as changes in the coefficient of friction for

the system as a whole, due to modifications of the random arrangement which would exist in the absence of Coulomb forces. The first order effect depends on the factor

$$(\rho_1 - \rho_2)^2 = \rho_1^2 - 2\rho_1\rho_2 + \rho_2^2$$

The three terms in the expansion of this square correspond to systematic changes of cation-cation, cation-anion and anion-anion distances, respectively. The longer distances between like ions are compensated by the shorter distances between ions of opposite charge, but the compensation of the resultant effects is not complete except when the individual coefficients of diffusion are equal.

The reason for the second order effect is that the reciprocal distances between two ions of charge ^+e, say, are not decreased quite as much as the reciprocal distances between ions of charges ^+e, ^-e are increased. The mutual viscous drag varies inversely as the distance, so that the resultant over-all reduction of interionic distances reduces the total drag.

The predicted variation of the "mobility for diffusion" with the concentration is much smaller than that predicted and found for the mobility of an ion in an electric field. Moreover, the initial decrease of the mobility due to the first order electrophoretic effect will be offset at higher concentrations by the second order effect. It will take very good experimental technique to get positive evidence for these effects.

On the other hand, the thermodynamic factor

$$d\mu/d \log c = kT(1 + c(d \log f/dc)),$$

which is involved in the coefficient of diffusion, exhibits quite a marked variation and is responsible for an easily measurable difference between the coefficients of diffusion at finite concentrations and the limiting values computed from Nernst's formula.

Problems of diffusion which involve more than two kinds of ions are generally difficult enough to compute under the assumption that the individual coefficients of diffusion are constant. So as to test the hearts of computers, we shall now give formulas for the corrections which must be applied when the effects of Coulomb forces are taken into account. Under the circumstances, we shall not give elaborate directions for the various applications, but only state the results in compact form.

The left member of Equation (14) is given in terms of the concentration gradients by Equation (41) or whatever improved formula may apply to the case in hand. Unfortunately, we must, in general, add the forces due to an electric field (diffusion potential) which is deter-

mined *implicitly* by the condition that no space charges shall accumulate from the motion of the ions.

It remains to specify the dissipation-function, which we shall write in the general form

$$2F(\mathbf{J}, \mathbf{J}) = \sum_{j,i} R_{ji}(\mathbf{J}_j \cdot \mathbf{J}_i) = \sum_j (\rho_j/c_j)\mathbf{J}_j^2 + 2F' + 2F'' \qquad (54)$$

The last two terms represent the corrections due to electrophoresis and relaxation forces, respectively Equation (44) gives us the (first order) electrophoretic effect

$$2F' = (2/3\eta\kappa\epsilon kT) \mid \sum e_j\rho_j\mathbf{J}_j \mid^2 \qquad (54a)$$

For the relaxation term $2F''$, we shall prefer a description which is simpler and more explicit than that given previously by Onsager and Fuoss. Our notation will differ a little from theirs; in particular t_1, t_2, ... will carry their usual connotation of electrolytic transference numbers, as follows

$$\begin{aligned}
\omega_i &= 1/\rho_i \\
\lambda_0 &= \sum c_i e_i^2 \omega_i \\
t_i &= c_i e_i^2 \omega_i/\lambda_0 = \Gamma_i\omega_i/\lambda_0 \\
\Gamma &= \sum \Gamma_i = \sum c_i e_i^2
\end{aligned}$$

We next define the quantities

$$0 = \alpha_1 < \alpha_2 < \cdots < \alpha_s$$
$$g_1, g_2, \cdots, g_s$$

implicitly by the identity in Θ

$$\left(1 - \sum_{i=1}^{s} \frac{t_i\omega_i^2}{\omega_i^2 - \Theta}\right)\left(1 + \sum_{p=1}^{s} \frac{g_p}{\alpha_p^2 - \Theta}\right) \equiv 1;$$

thus $\sigma_2^2, \ldots, \alpha_s^2$ are the roots of an equation of the order $s - 1$, given directly by a sum of partial fractions with positive numerators. Another form of that equation is

$$d(\alpha_p) = d(-\alpha_p) = q_p,$$

whereby

$$d(\zeta) = (\lambda/\Gamma) \sum_i t_i/(\omega_i + \zeta) \qquad (55)$$

Onsager and Fuoss defined $q_1, \ldots, q_s$ directly as the roots of a secular equation. The formulas given above are more convenient for compu-

tation. Moreover, the part of the dissipation-function due to the relaxation effect can be expressed in the relatively simple form

$$2F'' = \frac{\kappa}{3\lambda_0 \epsilon kT} \sum_{p=2} g_p (1 - q_p^{1/} \quad \sum_{i=1}^{s} \frac{e_i{}^{\prime}\omega_i}{\omega_i{}^2 - a} , \mathbf{J}_i \Big|^2 \tag{54b}$$

In some special cases, we get simple and completely explicit results.

For the case

$$\omega_1 = \omega_2 = \cdots = \omega_s$$

we may refer to the work of Onsager and Fuoss.[24]

One question which occurs now and then concerns the diffusion of one kind of ions present in small concentration in an electrolyte of otherwise nearly constant composition. In such a case, the activity coefficient is sensibly constant, the diffusion potential is small (by any reasonable convention), the electrophoretic effect is negligible and the relaxation effect can be computed explicitly. Thus, we get the following simple "limiting law" for the *coefficient of diffusion of an ion whose transference number is very small*

$$D_j = \omega_j [kT - (\kappa e_j^2 / 3\epsilon)(1 - (d(\omega_j))^{1/2})], \tag{56}$$

where $d(\zeta)$ is the function defined by Equation (55) above.

The connection with the previous results of Onsager and Fuoss[24] may be indicated briefly. Their formulas involve a matrix H whose elements h_{ji} can be described as follows[27]

$$h_{ji} - d(\omega_j)\delta_{ji} = (\lambda/\Gamma)t_i/(\omega_j + \omega_i)$$

The new procedure depends on the identity

$$\sum_{i=1}^{s} h_{ji} \left(\frac{1}{\omega_i + \zeta} + \frac{1}{\omega_i - \zeta} \right) = \frac{d(\zeta)}{\omega_j - \zeta} + \frac{d(-\zeta)}{\omega_j + \zeta}$$

which can be verified without difficulty by expansion in partial fractions. It is easily seen that each solution of the equation

$$d(\zeta) = d(-\zeta)$$

furnishes a solution for the eigenwertproblem of the matrix H. The rest of the computation is standard technique.

Concentrated Solutions of Electrolytes

Not much is known about the diffusion of concentrated electrolytes. We may reasonably expect a correlation between the viscosity and the coefficients of the dissipation-function, and we might even find some

[27] L. c.: 2760, equation (4. 7. 6).

simple relation like that described by Equation (34). It would seem that a relation simpler than this could only occur as a result of some accidental resemblance between solute and solvent.

For the rest, it is interesting to contemplate another possible effect which ought *not* to interfere with diffusion, although it seems to contribute much to the electrical resistance.

For concentration greater than one mol/l., the equivalent conductivities of strong acids in aqueous solution decrease sharply with increasing concentration. The absolute conductivities reach maxima which are approximately equal for hydrochloric, nitric and sulfuric acids; the conductivities of hydrochloric and nitric acids are closely comparable throughout. The variations of the viscosity are much too small to account for the resistance minima.

A likely explanation is that the passage of the ionic current leaves the water molecules in orientations opposed to that of the electric field and therefore unfavorable to the passage of further current. It is easy to see how the migration of a hydrogen ion by successive proton jumps must produce a result like that. It is not the same proton that makes the next jump! However, the approximate constancy of the transference numbers in HCl[28] seems to indicate that the anion is involved too.

If this interpretation is correct, then the observed value of the general resistance minimum

$$1.3 \text{ ohm cm} = 1.45 \times 10^{-12} \text{ sec}$$

is a measure for the dielectric relaxation time of water. Moreover, *the effect ought not to manifest itself in diffusion.* The evidence on hand is in accord with this contention[29]; a more extensive test is desirable.

[28] **Harned, H. S., & E. C. Dreby.** Jour. Am. Chem. Soc. **61:** 3113. 1939.
[29] **Gordon, A. R.** Jour. Chem. Phys. **5:** 522. 1937. **James, W. A., & A. R. Gordon.** Jour. Chem. Phys. **7:** 963. 1939.

A CONDUCTANCE METHOD FOR THE DETERMINATION OF THE DIFFUSION COEFFICIENTS OF ELECTROLYTES

By Herbert S. Harned and Douglas M. French

From the Department of Chemistry of Yale University, New Haven, Conn.

The possibility of employing conductance measurements for determining the diffusion coefficients of electrolytes has obviously occurred to many physical chemists, but this method has never been carefully exploited nor extensively used. As early as 1892, Niemöller[1] measured the change in conductance through a capillary tube containing a solution of a diffusing electrolyte. Haskell[2] measured the conductance at various heights of an electrolyte diffusing in a tube 50 cm. long and 5 cm. wide. More recently, Lamm[3] has employed conductance to determine diffusion constants of electrolytes in dilute solutions by a method which differs somewhat in principle from the one which we shall describe. In our preference for utilizing measurements in the later stages of the diffusion process we agree with Lemonde[4] whose procedure in other respects bears little resemblance to ours.

The most important characteristic of the method developed in this study consists in measuring vertical diffusion in a closed cell, by means of conductance measurements, between pairs of electrodes at suitable positions near the top and bottom of the cell. The idea of utilizing the difference of conductances between the bottom and the top of the cell was suggested by Professor Lars Onsager, some years ago, at which time he developed the theory of the method and computed the dimensions of a cell suitable for measurements in dilute solutions. A few preliminary experiments were made at Yale by Drs. Gosta Akerlof and Oliver A. Short which indicated promise of success. But difficulties arose due to control of temperature and convection, and to cell construction, which, in large part, have been eliminated as a result of this study.

THEORY OF CELL AND MEASUREMENT

The simplest form of cell for an electrolyte diffusing vertically upwards is a rectangular parallelopiped of height, a, and with electrodes

[1] **Niemöller, A.** Ann. der Phys. und Chem. **47**: 694. 1892.
[2] **Haskell, R.** Phys. Rev. **27**: 145. 1908.
[3] **Lamm, O.** Svensk Kem. Tidskrift **51**: 139. 1939; **55**: 263. 1943.
[4] **Lemonde, H.** Ann. de physique (11) **9**: 539. 1938.

at positions which may be most suitably determined by theory. Its schematic cross section is shown in FIGURE 1, where the electrodes are placed at a distance, ξ, from top and bottom. The cell is completely filled with the diffusing solution and measurements of conductance made as the process proceeds. Fick's law for the flow, $\mathbf{J}$, is

$$\mathbf{J} = c\mathbf{v} = -\mathfrak{D}\nabla c = -\mathfrak{M}\nabla\mu, \tag{1}$$

where c is the concentration per unit volume of diffusing component, μ its chemical potential, $\mathbf{v}$ its velocity, $\mathfrak{D}$ the diffusion coefficient and ∇

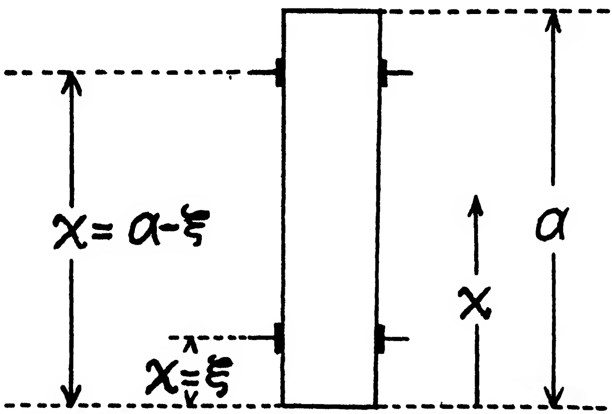


FIGURE 1. Vertical cross-section of cell showing quantities used in theoretical derivations.

the operator "del." Thus,

$$\mathfrak{D} = \left(\frac{\partial\mu}{\partial c}\right)_{P,T}\mathfrak{M}. \tag{2}$$

For the ideal solution,

$$\frac{\partial\mu}{\partial c} = \frac{RT}{c}, \qquad \text{whence} \qquad RT\mathfrak{M} = c\mathfrak{D}, \tag{3}$$

and, for the real solution,

$$\frac{\partial\mu}{\partial c} = \frac{RT}{c}\left[1 + c\,\frac{\partial \ln y_\pm}{\partial c}\right], \tag{4}$$

and

$$c\mathfrak{D} = RT\mathfrak{M}\left[1 + c\,\frac{\partial \ln y_\pm}{\partial c}\right], \tag{5}$$

where $y_\pm$ is an activity coefficient on a suitable concentration scale.

From the equation of continuity and Equation (1) we obtain

$$\frac{\delta c}{\delta t} = -\nabla \cdot (c\mathbf{v}) = \nabla \cdot \mathfrak{D}\nabla c = \frac{\partial}{\partial x}\left(\mathfrak{D}\frac{\partial c}{\partial x}\right). \tag{6}$$

Here, $\mathfrak{D}$ is the differential diffusion coefficient. As the experiment progresses, the concentrations at the top and bottom of the cell approach each other. We shall assume that this difference in concentration is sufficiently small so that $\mathfrak{D}$ may be taken to be constant. In which case

$$\frac{\delta c}{\delta t} = \mathfrak{D}\frac{d^2c}{dx^2} \tag{7}$$

with boundary conditions

$$\frac{\delta c}{\delta x} = 0 \tag{8}$$

for $x = a$, and $x = 0$.

The solution of Equation (7) which satisfies these boundary conditions[5] may be arranged in a Fourier series

$$c = \sum_{n=1}^{\infty} A_n e^{-\frac{n^2\pi^2\mathfrak{D}}{a^2}t} \cos\frac{n\pi x}{a} + c_0, \tag{9}$$

where the Fourier coefficients are such as to satisfy the initial conditions of concentration. The difference in concentration of electrolyte at the bottom (ξ) and top ($a - \xi$) electrodes is given by

$$\begin{aligned} c(\xi) - c(a - \xi) = {} & 2A_1 e^{-\frac{\pi^2\mathfrak{D}}{a^2}t} \cos\frac{\pi\xi}{a} \\ & + 2A_3 e^{-\frac{9\pi^2\mathfrak{D}}{a^2}t} \cos\frac{3\pi\xi}{a} \\ & + 2A_5 e^{-\frac{25\pi^2\mathfrak{D}}{a^2}t} \cos\frac{5\pi\xi}{a} \\ & + \cdots \end{aligned} \tag{10}$$

We note immediately that all even terms vanish and that, from the character of the exponential term the series converges very rapidly. If, as in our latest design of cell, $\xi = a/6$, then

$$c(\xi) - c(a - \xi) = 2A_1 e^{-\frac{\pi^2\mathfrak{D}}{a^2}t} + 2A_5 e^{-\frac{25\pi^2\mathfrak{D}}{a^2}t} + \cdots \tag{11}$$

and only the first term has significance after a sufficient time has

[5] Houstoun, R. A. An Introduction to Mathematical Physics: 88. Longmans, Green and Co. London. 1912.

elapsed. Since ξ and a have fixed values, the simple first order equation

$$\ln\,[c(\xi) - c(a - \xi)] = -\frac{\pi^2 \mathfrak{D}}{a^2}\,t + \text{constant} \tag{12}$$

is derived.

In order to render the numerical computations less complicated, we shall let K_B and K_T† equal the reciprocals of the resistances measured at the bottom and top electrodes of the cell, and make the further assumption that, in the range of concentrations measured $(c(\xi) - c(a - \xi))$ is proportional to $(K_B - K_T)$. This matter can be tested easily from the known conductance of the electrolyte under examination Equation (12) then reduces to

$$\ln\,(K_B - K_T) = -\,t/\tau + \text{constant}, \tag{13}$$

where $1/\tau = \pi^2 \mathfrak{D}/a^2$. Therefore, the slope of the line found by plotting $\ln(K_B - K_T)$ against t is $-\,1/\tau$, and the diffusion coefficient is given by

$$\mathfrak{D} = \frac{a^2}{\pi^2}\,\frac{1}{\tau}. \tag{14}$$

The great simplicity of the method is at once apparent, since a measurement of the depth of the cell, a, and the top and bottom conductances at suitable time intervals are the only data required to determine the diffusion coefficient.

At this juncture, it is well to examine the assumptions made in the above derivation which indicates that, as t increases, a plot of $\ln(K_B - K_T)$ versus t should be a straight line. As will be shown, this is actually true over quite a range of concentration differences Since $\mathfrak{D}$ is determined from the slope, our use of Equation (7) rather than Equation (6) is justified and the method yields a measure of the differential diffusion coefficient.

The reason for the experimental verification of a linear plot over quite a range in concentrations resides in employing two pairs of electrodes. Not only is this differential method responsible for the elimination of the disturbing terms in the Fourier series but also it helps to eradicate other effects by cancellation.

For example, let us test the assumption that $[c(\xi) - c(a - \xi)]$ is proportional to $(K_B - K_T)$ for solutions in the range of concentrations which we shall employ in our measurements. The specific con-

† K_B and K_T are values corrected for the slight difference in cell constants at the bottom and top pairs of electrodes. The uncorrected conductances (TABLE 2) are denoted $K_B{}^*$ and $K_T{}^*$, respectively.

ductance, L, may be computed by a series of the type

$$L = Ac + Bc^{3/2} + Dc^2 + Ec^{5/2} + \cdots \qquad (15)$$

and the difference of conductance at two concentrations will be given by

$$\begin{aligned} L_1 - L_2 &= A(c_1 - c_2) + B(c_1^{3/2} - c_2^{3/2}) \\ &\quad + D(c_1^2 - c_2^2) + E(c_1^{5/2} - c_2^{5/2}) + \cdots \\ &= A(c_1 - c_2) + B(c_1^{3/2} - c_2^{3/2}) \\ &\quad + D(c_1 - c_2)(c_1 + c_2) + E(c_1^{5/2} - c_2^{5/2}) + \cdots \qquad (16) \end{aligned}$$

Under the conditions of the experiment, the sum of the concentrations, $[c_1(\xi) + c_2(a - \xi)]$, is constant during the experiment. Therefore $L_1 - L_2$ is linear in $(c_1 - c_2)$, in the first and third terms, and the deviation from linearity in the second and fourth should not be great. In any case, $L_1 - L_2$ is much more nearly linear to $(c_1 - c_2)$ than L is to c.

This conclusion may be verified from the known conductances of potassium chloride solutions from the data of Shedlovsky, Brown and MacInnes[6] whose equation for the molecular conductance at 25° is

$$\frac{\Lambda + 59.79\sqrt{c}}{1 - 0.2274\sqrt{c}} = 149.86 + 141.9c + 29.24c \log c - 180.6c^2 \quad (17)$$

In TABLE 1, are given the data at concentrations covering the range used in the diffusion measurements. Note that, at the beginning of the

TABLE 1

EQUIVALENT AND SPECIFIC CONDUCTANCES OF POTASSIUM CHLORIDE AT 25°

c_1	Λ (Eq. 17)	$(L = \Lambda c/1000) \times 10^4$	$(L_1 - L_2)/(c_1 - c_2)$
0.0054	143.358	7.741332	0.142044
.0050	143.585	7.179250	.142005
.0046	143.822	6.615812	.141975
.0042	144.074	6.051108	.141962
.0040	144.203	5.768120	.141951
.0038	144.337	5.484806	.141949
c_2			
0.0018	145.986	2.627748	
.0022	145.596	3.203112	
.0026	145.243	3.776318	
.0030	144.919	4.347570	
.0032	144.766	4.632512	
.0034	144.618	4.917012	
c_0			
0.0036	144.476	5.201136	

[6] Shedlovsky, T., A. S. Brown & D. A. MacInnes. Trans. Electrochem. Soc. 66: 165. 1934.

measurement, the concentrations at bottom and top of cell would correspond to 0.0054 and 0.0018, respectively, and, as the diffusion proceeds, these approach the mean final concentration of c_0. It was necessary to compute Λ to three decimal places to insure a consistent series of slopes. It is clear from the table that the variation of $(L_1 - L_2)/(c_1 - c_2)$ is much smaller than the corresponding variation of Λ which clearly shows the advantage of using two pairs rather than one pair of electrodes. Indeed, the variation of $(L_1 - L_2)/(c_1 - c_2)$ is small enough to permit us to assume proportionality of $[c(\xi) - c(a - \xi)]$ to $(K_B - K_T)$ without further corrections.

However, another factor is present in the assumption that the measured conductances are proportional to the concentrations at the heights of the electrodes during diffusion. Although the cells (top and bottom) possess "geometrical cell constants" when a solution of uniform concentration lies between them, the distribution of electrolyte is non-homogenous during diffusion. In every case, the concentration beneath the electrodes is greater than that above, so that the majority of the lines of force are underneath the plane through the middle of the electrodes. As the diffusion proceeds, the ratio of the number of lines above this plane to the number below tends to increase. This condition is compensated for in our cell, since the same effect occurs between both the top and bottom electrodes, whereas it would be completely uncompensated for in a cell containing one pair of electrodes.

GENERAL EXPERIMENTAL ARRANGEMENTS

The apparatus consisted of the diffusion cell, placed in a large dessicator which was completely immersed in a water thermostat. The leads from the cell electrodes passed through glass tubes in a rubber stopper at the top of the dessicator. This was the best arrangement available for insuring constancy of temperature in the cell since the latter could not be placed in a liquid. The cell was mounted on a large piece of iron in the bottom of the dessicator. Besides increasing the heat capacity of this part of the system, the additional weight served to steady the apparatus. The fluctuations of temperature in the cell under these conditions must be very small after thermal equilibrium is reached, since the cell liquid was contained in the heat-insulating material, lucite, and the cell was surrounded by air within the insulating glass of the dessicator. The temperature control of the water thermostat was $\pm$ 0.15°.

The conductance measurements were made with an A.C. bridge, containing compensating capacitance for the cell and a Wagner ground.

All parts were carefully calibrated and lead corrections made. The accuracy of this bridge was more than sufficient, since, at the present stage of development of the method, the largest errors result from cell design and manipulation.

THE DIFFUSION CELL*

The successful operation of the cell depends primarily on filling the upper and lower portions with water and the salt solution, or two solutions at different densities, and bringing them into contact without producing convection currents. If these are produced, they probably do not cease until diffusion is complete. To overcome this difficulty, a sliding mechanism like the one in a Tiselius cell is employed. This is shown by the vertical cross-section of the cell in FIGURE 2 and the photographs of the filling and operating positions of the cell in PLATES 2 and 3, respectively. Lucite was found to be a satisfactory material for construction, since it could be machined very accurately.

The upper and lower parallelopiped sections were made equal in size The upper section could be readily moved from the filling (PLATE 2) to the diffusing position (PLATE 3). All flat surfaces were carefully lapped and the joints sealed with a rubberized stop-cock grease. This was found sufficient to prevent leakage.† The sliding surfaces were greased with vaseline. The electrodes were prepared by soldering a strip of sheet platinum on to a block of copper which was machined to the dimensions given in FIGURE 2.

EXPERIMENTAL TECHNIQUE

The lucite pieces of the cell were thoroughly cleaned with soap and water, then rinsed with water and allowed to dry. They were then screwed together after rubberized stop-cock grease was placed on all fixed adjacent surfaces. The surfaces facing the inside of the cell were then rendered plane by rubbing on pieces of fine emery paper and polishing cloth which rested on a flat plate. Removal of any grease was effected by rubbing with a cloth dampened with petroleum ether.

The electrodes were then platinized. Rather heavy black coatings were required to prevent polarization, since the electrodes were small and close together and the cell was not ideal for conductance measure-

* The cell described below is the one used for the measurements reported Note that the centers of the electrodes are not at distances from top and bottom equal to exactly 1/6 the depth of the cell. Equation (10), not (11) is therefore applicable. After sufficient time, only the first term on the right proves to be significant

† In one cell, lucite cement was tried without success Either this cement was applied incorrectly or the surfaces were not flat enough for successful application since the cell would not remain leak-proof.

ments. The cell sections were then washed in a stream of distilled water to remove all electrolytes, dried and assembled in the filling position and the appropriate solutions admitted into the upper and lower sections through the pair of ¼″ holes in the covers above them. Usually, conductivity water was placed in the upper section and salt solution in the lower.

In order to prevent the formation of any bubbles within the cell during the time (two or three weeks) in which measurements were taken, the following procedure was adopted. The cell was placed in the dessi-

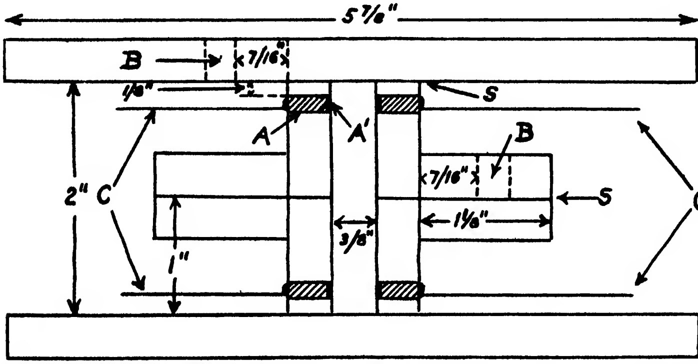


FIGURE 2 Vertical cross-section of cell showing essential parts. Unshaded parts are of lucite of [illegible] thickness

A—copper block 0 868″ × ⅛″ × 1½″
A′—0 007″ platinum foil soldered upon copper block
B—filling holes of ¼″ diameter
C—leads to electrodes
S—sliding surfaces for the top section

cator and the air exhausted. After a few hours, carbon dioxide free air was admitted and the dessicator was immersed in the thermostat. The leads from the electrodes passed through four glass tubes in the rubber stopper at the top of the dessicator. These were made air-tight by use of Apiezon sealing compound.

To insure thermal equilibrium, the apparatus was kept in the thermostat for 36 hours or more before commencing the diffusion. At the end of this time, the apparatus was removed momentarily from the thermostat, and the top section was slipped to the final position directly over the salt solution. In this way, a sharp, convection-free boundary was formed. After approximately 36 hours of diffusion, the measurements were suitable for calculation of the diffusion coefficient.

These experiments were performed with aqueous solutions of potassium chloride ranging from 0.0025 to 0.005 M at the end of the diffusion.

This may be computed easily from the accurately known concentrations of the solutions originally introduced in the upper and lower sections of the cell, since the volumes of these sections were equal. When distilled water was introduced in the upper section, the final concentration was simply one-half the initial concentration in the lower half.

To convert from molality to normality, the equation

$$\frac{c}{m} = 0.9970 - 0.0284m + 0.0003m^2 \tag{18}$$

was employed.[7]

EXPERIMENTAL RESULTS AND CALCULATION OF DIFFUSION COEFFICIENT

Four determinations were completed. Of these, we record in TABLE 2 the data from one of the most consistent. Readings were taken after the first twenty-four hours of diffusion and, at intervals (usually, of

TABLE 2

OBSERVED AND DERIVED DATA FOR THE COMPUTATION OF THE DIFFUSION COEFFICIENT OF POTASSIUM CHLORIDE SOLUTIONS AT 25°. SALT CONCENTRATION AT COMPLETION OF DIFFUSION EQUALS 0.00351 N.

(1) $K_B^* \times 10^3$; (2) $K_T^* \times 10^3$; (3) $(K_B^* + K_T^*) \times 10^3$; (4) $(K_B^* - K_T^* - 0.000009) \times 10^4$; (5) $\ln (K_B^* - K_T^* - 0.000009) \times 10^4$; (6) Four hour interval differences between successive values in column (5).

t(secs)	(1)	(2)	(3)	(4)	(5)	(6)
144,000	1.65854	0.74437	2.40291	9.0617	2.20295	.10866
158,400	1.61207	.79110	2.40317	8.1197	2.09429	.10903
172,800	1.57020	.83297	2.40317	7.2823	1.98545	.10903
187,200	1.53224	.87023	2.40247	6.5301	1.87642	.10894
201,600	1.49799	.90338	2.40137	5.8561	1.76748	.10852(a)
230,400	1.44003	.95968	2.39971	4.7135	1.55043	.10820
244,800	1.41579	.96378	2.39957	4.2301	1.44223	.10815
259,200	1.39381	1.00516	2.39897	3.7965	1.33408	.10889
273,600	1.37376	1.02428	2.39804	3.4048	1.22519	.10848
288,000	1.35608	1.04160	2.39768	3.0643	1.11671	.10870(a)
316,800	1.32552	1.07073	2.39625	2.4579	.89931	.10827
331,200	1.31284	1.08327	2.39611	2.2057	.79104	.10893(b)
352,800	1.29609	1.09977	2.39586	1.8732	.52765	.10869(b)
374,400	1.28159	1.11345	2.39504	1.5912	.46461	.10868(a)
403,200	1.26600	1.12895	2.39496	1.2805	.24725	.10904(a)
432,200	1.25346	1.14149	2.39494	1.0296	.02917	.10907(c)
457,200	1.24426	1.15019	2.39445	.8507	−.16170	.10899(d)
489,600	1.23478	1.15921	2.39399	.6657	−.40692	.10972(c)
514,800	1.22880	1.16486	2.39386	.5494	−.59893	

(a) Eight hour interval; (b) six hour interval; (c) seven hour interval; (d) nine hour interval.

[7] Harned, H. S., & B. B. Owen. The Physical Chemistry of Electrolytic Solutions. 556. Reinhold Publishing Corporation. New York. 1943.

four hours), for six days. Finally, the cell was rocked and placed on its side, inverted, and returned to its original position. These manipulations caused convection currents which produced a uniform concentration throughout the cell. In all these positions, measurements of the top and bottom resistances were made over a period of three days, until constant values of the resistances at top and bottom were obtained. Since the solution was now homogeneous, the ratio of these resistances was the cell constant ratio of the top and bottom pairs of electrodes. Successive readings of this ratio over two days were consistent to within 0.05%, and the difference in resistance between the top and bottom electrodes was 5.50 ohms.

Columns (1) and (2) contain the measured conductances (reciprocal resistances), at bottom and top of cell, corrected for leads and for the small differences in time taken between the readings of top and bottom electrode pairs. Column (3) shows that, under the conditions of the experiment, the sum of these conductances is nearly constant. This fact was utilized in the discussion of the assumption of proportionality of $(L_1 - L_2)$ to $(c_1 - c_2)$ immediately following Equation (16).

Column (4) contains the differences of the bottom and top conductances minus a small empirical constant selected so that the values of the successive differences of the logarithms of this quantity given in column (6) are as uniformly constant as possible in the intermediate range of time. This is equivalent to correcting for the cell constant and gives an independent check of its direct determination. For consider the equation

$$(K_B{}^* - K_T{}^*) - (K_B{}^\infty - K_T{}^\infty) = A'e^{-\frac{t}{\tau}} \tag{19}$$

where $K_B{}^\infty$ and $K_T{}^\infty$ are the designated conductances when t equals infinity. If k_B and k_T are the cell constants of the bottom and top cells, respectively, and k equals their ratio, k_T/k_B, then this equation becomes

$$(K_B{}^* - kK_T{}^*) - (K_B{}^\infty - kK_T{}^\infty) = \frac{A'}{k_B}e^{-\frac{t}{\tau}}, \tag{20}$$

or

$$(K_B{}^* - kK_T{}^*) = (K_B - K_T) = Ae^{-\frac{t}{\tau}}, \tag{21}$$

since the second term on the left of Equation (20) vanishes. Thus, plots of the logarithms of the left sides of Equations (19) and (21) versus t will have the same slope. At the end of the experiment, we found that $(R_T{}^\infty - R_B{}^\infty)$ was 5.50 ohms, and the average value of $R^\infty \cong 834$ ohms. Since $K = 1/R$, $\Delta K = K_T{}^\infty - K_B{}^\infty \cong -\Delta R^\infty/R^{\infty 2} = -7.9$

$\times 10^{-6}$ reciprocal ohms. This is in close agreement with -9×10^{-6} found from the diffusion data.

In column (6) of the table, we have subjected the data to a severe test. Here, the successive differences of logarithms for four hour intervals have been recorded. In FIGURE 3, these have been plotted against the number of the observation. In choosing a final value of 0.1086, we neglected the last four results, since towards the end of the experiment, $(K_B^* - K_T^*)$ becomes so small that a large error occurs in evaluating the slope. We note that, after forty hours of diffusion, all these results are within $\pm$ 0.0005 of the average value.

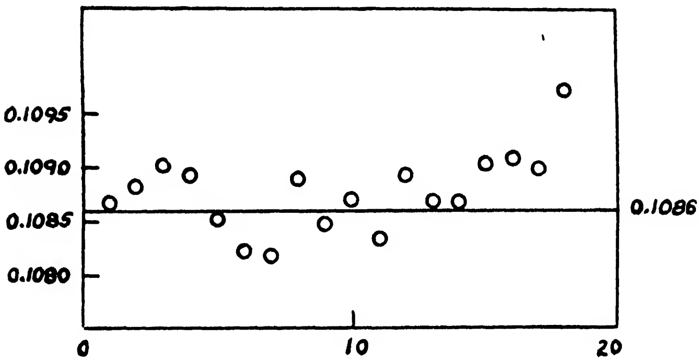


FIGURE 3. Plot of slopes given in the last column of TABLE 2.

From the value 0.1086, we find that the slope of the plot of $\ln(K_B - K_T)$ versus t is

$$-0.1086/14{,}400 = -7.542 \times 10^{-6}\ \text{sec.}^{-1} \qquad (22)$$

The cell height was 5.056 cm. Therefore, according to Equation (14)

$$\mathfrak{D} = \frac{a^2}{\pi^2}\frac{1}{\tau} = \frac{(5.056)^2}{\pi^2}\, 7.542 \times 10^{-6} = 1.953 \times 10^{-5}\ \text{cm}^2\ \text{sec}^{-1} \qquad (23)$$

at 0.00351 N.

Three other determinations of $\mathfrak{D}$ were made at 0.00250, 0.00348 and 0.00497 normal concentrations. The results are given in TABLE 4.

COMPARISON WITH THEORY

Onsager and Fuoss[8] have developed a theory for the variation of the diffusion coefficient of electrolytes with concentration. Their limiting

[8] Onsager, L., & R. M. Fuoss. J. Phys. Chem. 36: 2689. 1932.

law and the limiting value, $\mathfrak{D}_0$ of the diffusion coefficient may be expressed by the equations

$$\mathfrak{D} = \mathfrak{D}_0 - \mathfrak{S}_{(\mathfrak{D})} \sqrt{c} \tag{24}$$

$$\mathfrak{D}_0 = 17.863 \times 10^{-10}(\lambda_+^0\lambda_-^0/\Lambda^0)T \tag{25}$$

$$\mathfrak{S}_{(\mathfrak{D})} = \frac{3.732 \times 10^{-8}}{D_0^{1/2}T^{1/2}}\left(\frac{\lambda_+^0\lambda_-^0}{\Lambda^0}\right) + \frac{3.659 \times 10^{-8}}{\eta_0 D_0^{1/2}T^{-1/2}}\left(\frac{\lambda_+^0 - \lambda_-^0}{\Lambda^0}\right)^2 \tag{26}$$

where λ_+^0, λ_-^0 and Λ^0 are the equivalent ionic and molecular conductances at infinite dilution, D_0 the dielectric constant of water, η^0 its viscosity and $\mathfrak{S}_{(\mathfrak{D})}$ is the limiting slope.*

Introducing the values: $\lambda_+^0 = 73.48$, $\lambda_-^0 = 76.34$, $\Lambda^0 = 149.82$[9], $\eta_0 = 8.949 \times 10^{-3}$, and $D_0 = 78.54$, we find that for potassium chloride at 25°, $\mathfrak{D}_0 = 1.994 \times 10^{-5}$ and $\mathfrak{S}_{(\mathfrak{D})} = 1\ 162 \times 10^{-5}$.

At higher concentrations, the theoretical equation for the diffusion coefficient is

$$\mathfrak{D} = 16.632 \times 10^{10}T\left(\frac{\overline{\mathfrak{M}}}{c}\right)\left(1 + c\frac{\partial \ln y_\pm}{\partial c}\right), \tag{27}$$

where

$$\left(\frac{\overline{\mathfrak{M}}}{c}\right) \times 10^{20} = 1.074\frac{\lambda_+^0\lambda_-^0}{\Lambda^0} - \frac{22.00}{\eta_0(DT)^{1/2}}\left(\frac{\lambda_+^0 - \lambda_-^0}{\Lambda^0}\right)^2\frac{\sqrt{c}}{1 + A'\sqrt{c}} + \frac{9.18 \times 10^7}{\eta_0(DT)^2}c\varphi(A'\sqrt{c}). \tag{28}$$

The activity coefficient $y_\pm$ on the mols per liter scale is related to that on the mol fraction scale, $f_\pm$ by

$$\log y_\pm = \log f_\pm - \log\left\{\frac{d + 0.001c(2M_1 - M_2)}{d_0}\right\}, \tag{29}$$

where M_1 is molecular weight of solvent, M_2 the molecular weight of electrolyte, d is the density of the solution and d_0 the density of solvent. The Debye and Huckel equation

$$\log f_\pm = -\frac{\mathfrak{S}_{(f)}\sqrt{c}}{1 + A'\sqrt{c}} + Bc \tag{30}$$

expresses $f_\pm$ as a function of the concentration. From these last two equations, we find that

$$1 + c\frac{\partial \ln y_\pm}{\partial c} = 1 - \frac{1.1514\,\mathfrak{S}_{(f)}\sqrt{c}}{(1 + A'\sqrt{c})^2} + 4.606\,Bc - c\psi(d), \tag{31}$$

* The equations and symbols used here are those of Harned, H. S., & B. B. Owen. The Physical [illegible] of Electrolytic Solutions. 178–180. Reinhold Publishing Corporation. [illegible] 1943.

[9] Shedlovsky, [illegible] Am. Chem. Soc. 54: 1423. 1932. Longsworth, L. G., & D. A. MacInnes. [illegible] 2073. 1938.

where

$$\psi(d) = \frac{\dfrac{\partial d}{\partial c} + 0.001\,(2M_1 - M_2)}{d + 0.001c\,(2M_1 - M_2)}. \qquad (32)$$

In these equations, $A'\sqrt{c} = \kappa a = \kappa \mathring{a} \times 10^{-8}$, where a and $\mathring{a}$ represent the mean distance of approach of the ions in centimeters and Ångstrom units, respectively.

In the dilute concentration range, the calculation of $\mathfrak{B}$ is considerably simplified, since the terms $c\psi(d)$ in Equation (31) and the second term on the right of Equation (28) which involves the square of a small quantity are negligible. For potassium chloride at 25°, $\mathring{a} = 3.8$ and $B = 0.0202$.[10] Using these values, and the values of the conductances, viscosity and dielectric constant given above, Equation (28) reduces to

$$\left(\frac{\mathfrak{M}}{c}\right) \times 10^{20} = 40.212 + 18.71\, c\varphi(A'\sqrt{c}). \qquad (33)$$

By employing values of $\varphi(A'\sqrt{c})$ obtained from a plot of this function given by Harned and Owen,[11] and this equation, the results in the second column of TABLE 3, were obtained. The third column of this

TABLE 3

DIFFUSION COEFFICIENTS ACCORDING TO THEORY OF ONSAGER AND FUOSS

c	$\left(\frac{\mathfrak{M}}{c}\right) \times 10^{20}$	$\left(1 + c\frac{\partial \ln y_{\pm}}{\partial c}\right)$	$\mathfrak{B} \times 10^5$	$(\mathfrak{B}_0 - \mathfrak{S}_{(\mathfrak{B})}\sqrt{c}) \times 10^5$
0.0	—	—	1.994	1.994
.0005	40.235	0.9877	1.971	1.968
.001	40 250	.9830	1.962	1.957
.0025	40.288	.9743	1.947	1.936
.0035	40.309	.9702	1.939	1.925
.005	40.335	.9654	1.931	1.912
.01	40.415	.9543	1.913	1.878

table contains values of the activity coefficient function calculated by Equation (31). The theoretical diffusion coefficient evaluated by Equation (27) is recorded in the next to last column. The last column contains values computed by the limiting law.

A comparison of the observed results with the theoretical is made in TABLE 4. The agreement in all cases is within one per cent. Judging from the consistency of the observations and the cell constant corrections, these should not be weighted equally. The first and third results are probably the best, while the second and fourth are somewhat low.

[10] Harned, H. S., & B. B. Owen. loc. cit.: 381.
[11] Harned, H. S., & B. B. Owen. loc cit.: 130.

FURTHER DISCUSSION OF THE METHOD AND SOURCES OF ERROR

Although only four measurements have so far been obtained, they are sufficient to show some of the advantages of the method and to indicate that considerable improvement in accuracy may be achieved As previously stated, it is an absolute method for measuring the differential diffusion coefficient and does not require any solutions for the calibration of the cell. It is most adaptable for use at low concentrations. Indeed, our determinations are at concentrations of the order of one-tenth those previously employed for absolute measurements.

Even though initial convection currents are greatly reduced by using

TABLE 4

OBSERVED AND CALCULATED DIFFUSION COEFFICIENTS OF POTASSIUM CHLORIDE AT 25°

	$D \times 10^5$ (obs.)	$D \times 10^5$ (theo.)
0.0025	1.944	1.947
.00348	1.921	1.939
.00351	1.953	1.939
.00497	1.913	1.931

sliding surfaces to bring the solutions in contact initially, some convection will remain. A density gradient will reduce convection in time, but, as this gradient becomes less, convection, if present, will not be eliminated. This effect will be one factor in determining the lower limit of concentration at which successful measurements may be obtained.

Difficulty was encountered in maintaining water of high conductance in the upper half of the cell before diffusion was started. Indeed, in the experiment described, the water conductance was of the order of one per cent of the final conductance (K^∞) of the cell. The source of the ions which caused this increase has not been determined. Although this effect should be reduced to a minimum, it may have little influence on the observed diffusion coefficients at the salt concentrations employed. This is due to the fact that we use a difference in conductance, ($K_B - K_T$), and the conductance caused by the ions of water at the bottom and top electrodes may be largely eliminated by cancellation.

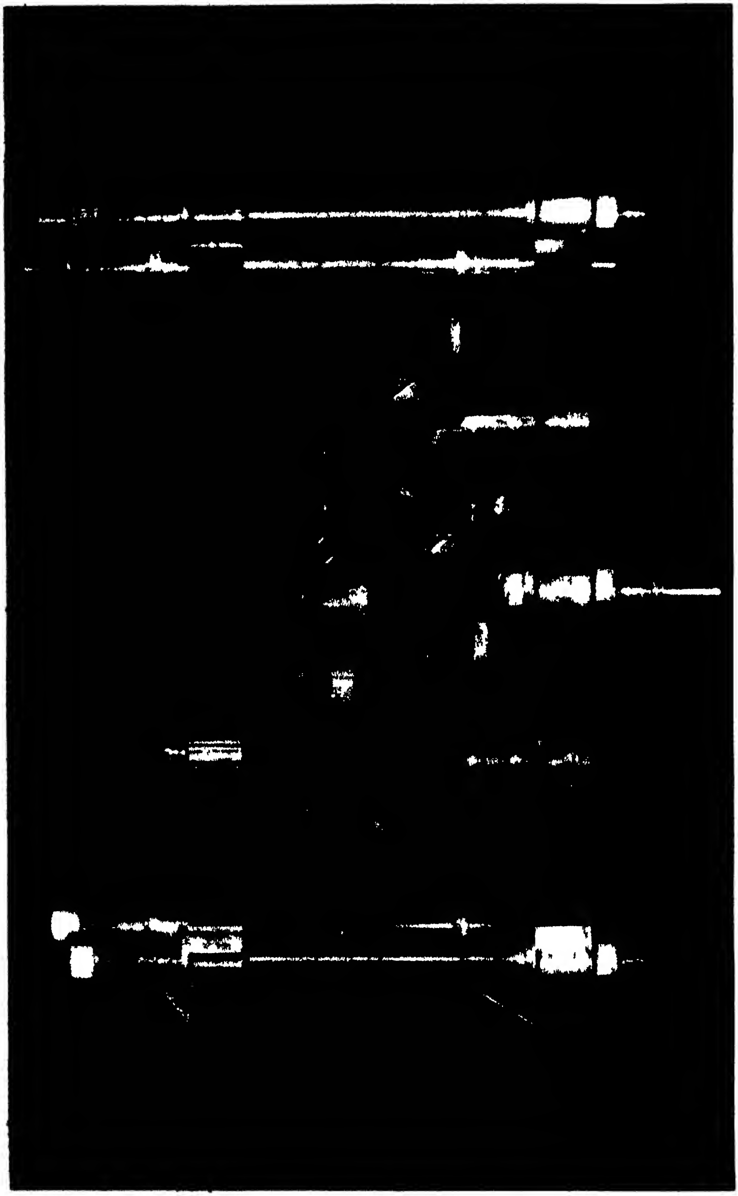
One of the potentially serious sources of error may be caused by carrying a little grease into the middle of the cell, when the sliding surfaces are brought into the diffusing position. This would slightly de-

crease the cross sectional area. To eliminate this effect completely, another cell has been designed with the sliding surface at the bottom. So far, we have acquired no data with this cell.

Vibration in the cell was probably not completely eliminated, but must have been very slight. The thermostat vibrated very little and the precaution of mounting the cell on two steel plates upon a rubber mat was taken.

SUMMARY

An apparatus is described for the determination of the diffusion coefficients of electrolytes by measuring the differences in conductance across the top and bottom of a cell of the form of a rectangular parallelopiped, while the electrolyte is diffusing vertically upward. The theory of the cell is developed, and both the theoretical and practical advantages of employing a difference in conductance rather than a single conductance measurement are shown. In addition to the conductance measurements at suitable time intervals, only the depth of the cell is required for the absolute determination of the diffusion coefficient. Measurements of the diffusion coefficient of potassium chloride solutions at 25° and in the concentration range 0.0025 to 0.005 molal have been made. These are compared with theoretical values derived from the theory of Onsager and Fuoss. The error of these first determinations is estimated to be approximately ±0.9%. Our experience indicates that considerable improvement in accuracy can be effected.



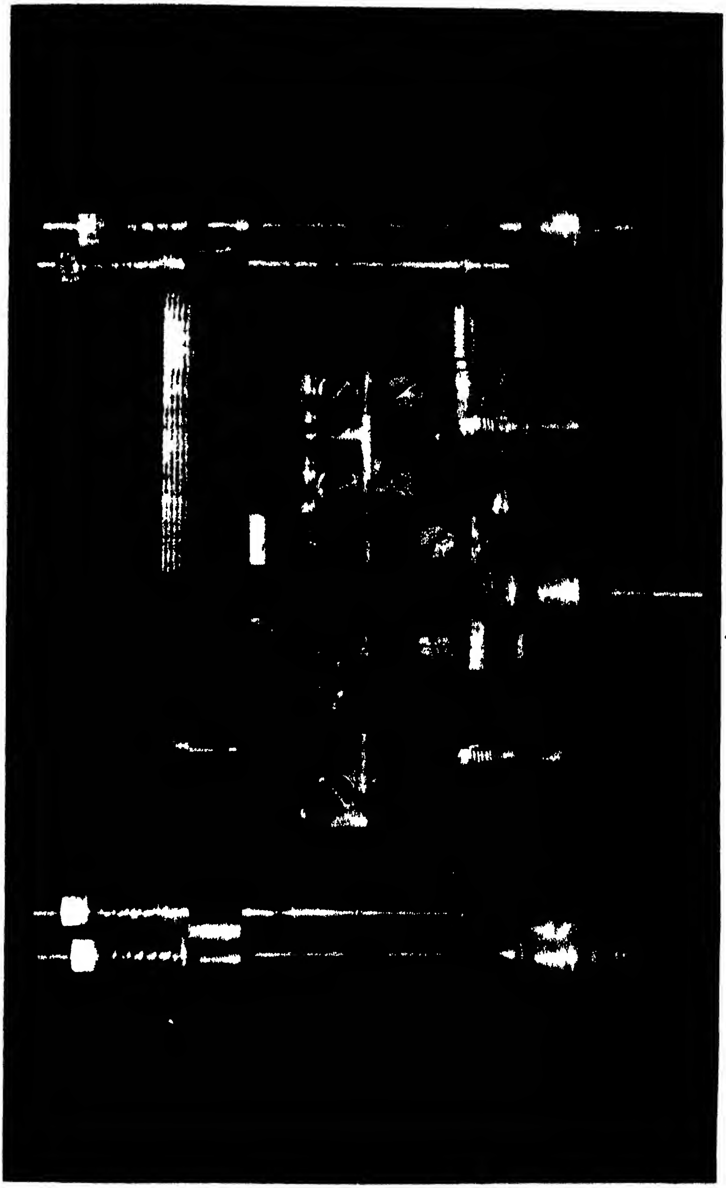
HARNED AND FRENCH · DETERMINATION OF DIFFUSION COEFFICIENTS

PLATE 2

Photograph of cell with foremost sliding guide removed. Cell is in filling position.

PLATE 3

Photograph of cell with foremost sliding guide removed. Cell is in diffusing position.



HARNED AND FRENCH · DETERMINATION OF DIFFUSION COEFFICIENTS

THE DIAPHRAGM CELL METHOD OF MEASURING DIFFUSION

By A. R. Gordon

Chemistry Department, University of Toronto, Toronto, Canada

INTRODUCTION

In the whole field of physical chemistry, there is probably no problem which has so exercised the ingenuity of experimenters as the determination of the diffusion coefficient. The reason lies in the nature of the diffusion process itself. Transport by diffusion is, at best, slow and, unless special precautions are taken, it may be masked more or less completely by thermal or mechanical convection. In 1928, Northrop and Anson [1] described a new type of diffusion apparatus which seemed, at first glance, to avoid most of the difficulties inherent in the earlier techniques. By confining the diffusion to the pores of a sintered diaphragm, they were able to work with relatively enormous concentration gradients, thus reducing greatly the time required for a measurement. Moreover, mechanical or thermal disturbances could no longer affect the diffusion column. Since that time, however, some awkward questions have been raised as to the validity of certain assumptions on which the original theory of the method was based. It is with this aspect of the diaphragm cell problem that the present paper is concerned.

In its simplest form, the cell is a bell-shaped glass vessel, fitted with a stop-cock at the top, and with the flat diaphragm of sintered glass or alundum forming the base. Various sizes have been used varying from 10 to 200 cc. capacity; those employed by McBain and Liu [2] and in my laboratory, as a rule, have been from 90 to 100 cc. with a diaphragm approximately 5 cm. in diameter and from 2 to 3 mm. thick. The effective diameter of the pores in the diaphragm must be such that gross streaming through the diaphragm is avoided, but, at the same time, the pores must be large in comparison with molecular dimensions so that the diffusion will take place under conditions comparable with those obtaining in free diffusion. The actual pore size to satisfy these conditions can vary considerably. With Jena glass diaphragms, the sizes most frequently used have been G-4 (pore diameter 2-5 microns)

[1] **Northrop, J. H., & M. L. Anson.** J. Gen. Physiol. **12:** 543. 1928.
[2] **McBain, J. W., & T. H. Liu.** J. Am. Chem. Soc. **53:** 59. 1931.

and G-3 (15-20 microns). McBain and Dawson,[3] in some experiments, used a G-5 diaphragm (pore diameter < 1 micron) but found no resulting advantage. Two different alundum diaphragms have been tested by McBain and Liu: Norton, RA 225 and RA 98, pore diameter 8-10 and 10-15 microns respectively. Most experimenters seem agreed, however, that the G-4 provides the most generally useful combination of characteristics.

In the original form of the apparatus, the cell contains the stronger solution (referred to in this paper as the inner solution) and is suspended with its diaphragm just touching the upper surface of the weaker (outer) solution, which is contained in a glass vessel slightly larger than the cell. The lower surface of the diaphragm must be horizontal so that it can touch the outer solution over its whole area, and must be as near the upper surface of the outer solution as possible. In this way, secondary diffusion into the annular dead space above the bottom of the diaphragm is minimized. For this reason, it is convenient to support the cell with an adjustable suspension so that its position can be fixed accurately. The whole apparatus is placed in an air-bath whose temperature is regulated to 0.01° C; temperature gradients from the bottom to the top of the apparatus particularly are to be avoided, and, in some of the later work in the Toronto laboratory, we found it advantageous to enclose each cell and its outer vessel in a heavy copper cylinder with a copper lid to serve as a thermal shield. Evaporation from the exposed surface of the outer solution must naturally be prevented, if the experiment is a long one; e.g., by a thin rubber sleeve connecting the top of the outer vessel with the cell, or by placing in the air bath a solution isotonic to the outer solution.

The cell is cleaned by repeatedly drawing through the diaphragm first the solvent and then the solution with which it is eventually to be filled. It is essential to remove all air from the pores during this process, and also to avoid any contamination of the diaphragm with stop-cock grease. Moreover, the solutions must be outgassed to prevent formation of bubbles during the experiment. When the cell is finally filled, it is placed in contact with the outer solution and diffusion is allowed to proceed until the concentrations in the diaphragm are those for a steady state; i.e., for a solute whose diffusion coefficient is independent of concentration until the concentration is linear across the diaphragm. A rough but safe rule for estimating the time necessary to do this is to set $Dt/l^2 = 1.2$, where D is the diffusion coefficient,

[3] McBain, J. W., & C. R. Dawson. Proc. Roy. Soc. A 148: 32. 1935

t the time, and l is the effective pore length of the diaphragm. The justification for this is that the effect of the initial constant concentration in the diaphragm tends to disappear with exp $(-\pi^2 Dt/l^2)$. Thus, for $l = 0.4$ cm (see section b) and $D = 2 \times 10^{-5}$ cm^2/sec., the time required is of the order of three hours. In some experiments in my laboratory, the preliminary diffusion was followed by a second but much shorter preliminary diffusion with a fresh sample of outer solution, but no significant difference was apparent in the final results.

When "steady state" concentrations have been attained in the diaphragm, the cell is placed in contact with the final sample of the outer solution, and the diffusion is allowed to proceed for a suitable length of time. Except when the inner solution is analyzed *in situ*, e.g., by conductance, as in Hartley and Runnicles' measurements,[4] the cell at the end of the experiment is removed from contact with the outer solution; part of the inner solution is forced through the diaphragm by air pressure, thus flushing the diaphragm, and discarded; the remainder of the inner solution is then forced out of the cell, and both solutions are analyzed.

The apparatus and technique just outlined correspond to the procedure of many experimenters, but there have been some modifications. For example, McBain and Dawson [3] used a cylindrical glass apparatus with the diaphragm in the center dividing it into two compartments. Each compartment was fitted with two stop-cocks, so that both could be emptied or filled without disturbing the liquid in the diaphragm. In this way, they were able to replace not only the weaker solution, but the stronger, as well, with fresh samples for the final diffusion. Mouquin and Cathcart [5] used a similar cell, but mounted it on an axle in the plane of the diaphragm, so that the cell could be rotated slowly end over end as the diffusion proceeded. Enclosed in each compartment, were two glass spheres, one slightly heavier and the other slightly lighter than the solution, whose movement as the cell rotated stirred the solutions. Hartley and Runnicles[4] achieved mechanical stirring in a somewhat different manner, since their cell was rotated about its longitudinal axis on an axle inclined slightly to the vertical. The upper compartment (holding in this case the weaker solution) contained a glass sphere whose density was such that it rested lightly on the diaphragm. The lower compartment contained a lighter sphere which floated against the bottom of the diaphragm. As the cell rotated, the spheres rolled on the diaphragm and thus stirred the solutions. The

---

[4] Hartley, G. S., & D. F. Runnicles. Proc. Roy. Soc. A 168: 401. 1938.
[5] Mouquin, H., & W. H. Cathcart. J. Am. Chem. Soc. 57: 1791. 1935.

highly important results obtained in cells with mechanical stirring will be discussed in detail in Section B (page 291).

The elementary theory of the method is due primarily to Northrop and Anson [1] and to McBain and Liu.[2] The diaphragm is considered to be equivalent to a collection of parallel pores of average effective length l and of total effective cross-sectional area A. Both solutions are assumed to be of uniform composition right up to the entrance of the pores, and the transport from one compartment to the other is assumed to take place solely by diffusion; i.e., streaming in the pores and surface transport along their walls are excluded. For large enough volumes of the inner and outer solutions, their concentrations will change so slowly that the distribution of concentration across the diaphragm will not differ appreciably from that for a steady state. Thus, if C' and C'' are the concentrations of the inner and outer solutions at some instant of the diffusion, the slope of concentration for a solute, whose diffusion coefficient D is independent of concentration, will be $(C' - C'')/l$. This will be constant from the upper surface of the diaphragm ($x = 0$) to the lower surface ($x = l$). Hence, the amount of solute dq diffusing in an interval dt will be given by

$$- dq = DA(C' - C'') \cdot dt/l \tag{1}$$

Thus, if A/l be known, a knowledge of dq/dt, C' and C'' will fix D. This is essentially the method of Northrop and Anson. Since neither A nor l can be measured directly, they determined their ratio by calibrating the cell with a solute of known diffusion coefficient. In this respect, a diaphragm cell measurement resembles the usual conductance determination and, consequently, suffers under the same disadvantage in that the results are no more accurate than the "standard" on which they are based (See Section E, page 297).

Although Northrop and Anson gave an integrated form of Equation (1), they used the differential form for their own results by permitting diffusion to proceed for such a short time that $(C' - C'')$ was sensibly constant throughout the experiment. McBain and Liu[2] pointed out the practical advantages of longer experiments and, accordingly, used the integrated form of the equation. Let V' and V'' be the volumes of the inner and outer solutions. Since the change in the amount of solute in a compartment, in a time dt, must equal the quantity which has diffused into or out of the compartment, it follows that

$$V'dC' + (DA/l) \cdot (C' - C'')dt = 0 \tag{2a}$$

$$V''dC'' + (DA/l) \cdot (C'' - C')dt = 0 \tag{2b}$$

where dC' and dC'' are the changes in concentration in the interval dt.

Combining (2a) and (2b), there results

$$d(\Delta C)/(\Delta C) + \beta D dt = 0 \tag{3}$$

where ΔC stands for the concentration difference $(C' - C'')$, and β, the "cell factor," is given by $\beta = (A/l)\cdot(1/V' + 1/V'')$. Since it has been assumed that D is independent of C, it follows that Equation (3) can be integrated from the start of the experiment when the concentrations have their initial values C_0' and C_0'', to the end when they have attained their final values C_f' and C_f'':

$$\ln \Delta C_f / \Delta C_0 = -\beta D t \tag{4}$$

Equation (4), or some form equivalent to it, is usually employed to compute the diffusion coefficient in a diaphragm cell measurement. It should also be noted that, in effecting the integration, it has been assumed implicitly that the cell factor remains constant throughout the experiment, i.e., that both solutions are constant in volume.

The remainder of the paper will be devoted to a consideration of the validity of Equation (4) in interpreting diaphragm cell data.

A. THE QUESTION OF CELL VOLUME

At first glance, the definition of the cell factor suggests that, as long as the volume of each solution is the same in the calibration and in the subsequent measurements, the actual volume is a matter of indifference. This is, in fact, the case, provided all four concentrations in the ratio on the left of Equation (4) are known, as in McBain and Dawson's measurements.[8] In many instances, however, only the outer solution is replaced by a fresh sample after the preliminary diffusion, so that the initial concentration of the inner solution must be computed by taking advantage of the fact that

$$V'C' + V''C'' + V'''(C' + C'')/2 \tag{5}$$

is constant throughout the experiment, where V''' is the volume of the pores in the diaphragm. It should be noted that it is only when $V' = V''$ that the sum $(C' + C'')$ remains constant, and that the last item in Equation (5) can be ignored.

The usual method[2] of determining the cell and diaphragm volumes is to weigh the cell, draw in by suction sufficient liquid of known density to fill the diaphragm; and weigh again. Additional liquid is then drawn in to fill the cell, which is again weighed. In spite of its apparent crudity, this method gives surprisingly reproducible results, under optimum conditions, to a tenth of a per cent or so in the cell volume. Actually, the uncertainty in the measured diffusion coefficient due to

error in the volume calibration is, in practice, entirely negligible. If we write Equation (4) in terms of the experimentally determined quantities, it becomes

$$-\beta Dt = \ln \frac{1 - j}{1 + j(1 + r)} \tag{4a}$$

where j stands for $(C_f'' - C_o'')/(C_f' - C_o'')$ and $(1 + r)$ for the ratio $(V'' + V'''/2)/(V' + V'''/2)$. Note that j is, in effect, the ratio of the final concentrations of the two solutions counting the initial outer solution as solvent. Suppose, for purposes of illustration, that it had been intended to use equal volumes for the two solutions, but that the assumed cell volume V' is actually 1 per cent too great; this is equivalent, closely enough, to $r = 0.01$. Equation (4a) can then be written approximately

$$-\beta Dt = \ln (1 - j)/(1 + j) - 0.01j/(1 + j) \tag{4b}$$

the last item in Equation (4b) representing the error introduced by the incorrect cell volume. If, in the calibration, the diffusion is allowed to proceed until the difference in concentration has fallen to 40 per cent of its initial value, j is 3/7; the leading term on the right of Equation (4b) will then have the value -0.916; and the error in the cell factor will be 0.33%. If, however, in a subsequent measurement, j is also 3/7, the percentage error in βDt will again be 0.33%, and, consequently, will be allowed for automatically in the cell factor.

Even if the conditions vary considerably in the later measurements from those obtaining for the calibration, the resultant error is not serious. For example, suppose, with the cell just discussed, j is 1/4 in a subsequent measurement, i.e., $\Delta C_f/\Delta C_o = 0.6$; the percentage error in βDt will now be 0.39%, but the net resultant error in the measured diffusion constant will be only 0.06%.

On the other hand, a random (in contrast to a consistent) error of 1% in the volume of the outer solution is obviously a much more serious matter, since the resultant error in $\Delta C_f/\Delta C_o$ will not be absorbed in the cell factor, and, moreover, the true cell factor for the run will be different from that obtained in the calibration. For example, suppose, in the calibration, equal volumes for the two solutions are used, but that, in a later measurement, the volume of the outer solution is 1% too great; the resultant error in the computed $-\beta Dt$ will be approximately

$$-0.01j/(1 + j) + 0.005 \ln (1 - j)/(1 + j)$$

i.e., for $j = 3/7$ (as above) will be 0.8%. In brief, here, as in every relative measurement, consistency is all important, and it is for this

reason that it is advisable to use special burettes calibrated to deliver the outer solutions for each of the cells.

In the integration of Equation (4), there is another small volume error which has generally been ignored. In a typical diaphragm cell measurement, the volume of the inner solution is fixed, and this is in conformity with the usual definition of the diffusion coefficient in terms of transport relative to a fixed column of solution.[6] The volume of the outer solution and, consequently, the cell factor, however, will only remain constant provided the specific volume is linear in the concentration over the entire range of concentrations involved in the experiment. In general, this condition is fulfilled closely enough when the concentration range is of the order of a few tenth's of a mole per liter, but, for wide ranges, a slight error may be introduced, which can, however, be readily estimated in any given case.

B. HOMOGENEITY OF THE INNER AND OUTER SOLUTIONS

The elementary theory of the diaphragm cell sketched above assumes that both solutions are of uniform composition throughout. In the absence of mechanical stirring, the usual argument advanced in favor of this is that, in both compartments, the least dense solution tends to be at the bottom, and, thus, the contents of both compartments are stirred automatically by density difference. While there can be no doubt that this is true of the great bulk of the liquid in each compartment, there remains the possibility that stagnant layers may exist immediately adjacent to the diaphragm, through which diffusion will proceed as well as in the pores themselves. The probability that such layers exist has been abundantly demonstrated by various kinetic studies of reactions occurring at a solid-liquid interface. The familiar electrochemical phenomenon of the "limiting current" is an example.

Mouquin and Cathcart [5] demonstrated this by calibrating their cell both at rest and when rotating. They found that the static cell factor was some 5% less than that obtained with mechanical stirring, the obvious explanation being that stagnant layers were present with density stirring, but were absent (or nearly so) with mechanical stirring. The question at once arises whether such a difference can be accounted for by layers of a reasonable thickness. For a typical G-4 cell, such as those in use in the Toronto laboratory, A/l (computed from the cell factor and the volumes) is of the order of 10 cm; the pore volume of the diaphragm (which is Al) is approximately 1.5 cc.; hence, l is about 4 mm., some 60%, greater than the apparent thickness of the diaphragm,

[6] The alternative definition in terms of transport relative to solvent has certain mathematical advantages (See reference [30]).

viz., 2.5 mm. If the decrease of 5% in cell factor be ascribed to a 5% increase in effective l with density stirring, this implies layers only 0.1 mm. thick, an entirely possible value.

The next question is whether such layers make a reproducible contribution, independent of the solution, to the effective l of the diaphragm. There is, fortunately, definite evidence on this point, supplied by measurements with dilute potassium chloride solutions by Hartley and Runnicles[4] and by McBain and Dawson,[3] to which may be added some unpublished data from my laboratory. Hartley and Runnicles used mechanical stirring in the cell previously described; while, in the other two series, density stirring was employed. Mehl and Schmidt[7] also made some measurements in this range, but unfortunately do not state their results with sufficient precision for purposes of comparison with the other data. It is sufficient to say that McBain and Dawson's results and my own, if brought to the calibration they use, agree with their diffusion coefficients for solutions between 0.01 N and 0.1 N within 1% or better. The results are summarized in TABLE 1. The first column gives the initial concentration of the stronger solution, the other experimental conditions[8] being indicated in the heading of the table. The quantity tabulated is D_t/D_0, the ratio of the diaphragm cell integral coefficient calculated by Equation (4) to the coefficient at infinite dilution. The reason for using this quantity, rather than D_t itself, is that the density stirred measurements were at 25° C, while Hartley and Runnicles' were at 18° C. While the diffusion coefficient itself is strongly temperature dependent, its percentage variation, with concentration for solutions as dilute as these, is practically unchanged[9] for a shift of 7°. Both series have been brought to a common calibration for purposes of comparison, the value in heavy type at the head of the columns.

In view of the importance of Hartley and Runnicles' mechanically stirred results, it is advisable to consider their measurements in some detail. They used a G-4 and also a coarser G-3 diaphragm. With the former, the measured βD_t was independent of the speed of rotation at all concentrations; with the latter, the same was true for solutions

[7] **Mehl, J. W., & C. L. A. Schmidt.** Univ. Calif. Publ. Physiol **8**: 165. 1937.

[8] **McBain & Dawson** do not give explicit information as to the duration of their experiments but the method of analysis they used required that the diffusion proceed for a considerable length of time It will be shown in Section F that, for these conditions, the integral coefficient is not very sensitive to the actual value of $\Delta C_f/\Delta C_0$.

[9] A calculation of the change in the thermodynamic factor $(1 + d \ln f/d \ln C)$, between 18° and 25°, shows that it changes at most by 0.04% the changes in the **Onsager-Fuoss** mobility term (reference [30]) and in the relative viscosity are even less important. In the calculation of the thermodynamic factor, the activity coefficients of **Harned, Janz & Gordon** were employed (J Am Chem. Soc **64**: 513. 1942).

0.01 N or weaker; but βD_t increased with speed of rotation for solutions stronger than this. They interpreted this as evidence that streaming through the diaphragm only occurred with the coarser diaphragm, and, even here, only if the density difference between the solutions was appreciable. For their most dilute experiments, they found a slight increase in βD_t for the G-4, as compared with the G-3, and they attributed this to contamination of the solutions with soluble impurities from the diaphragm, the effect being more pronounced with the finer diaphragm owing to the greater surface area of its pores. Their explanation would seem reasonable, since they used a conductometric analysis which would be particularly sensitive in very dilute solution to traces of alkali dissolved from the diaphragm.

Before comparing Hartley and Runnicles' Table III with the density stirred results, it is necessary to allow for certain differences in their technique from that in the density-stirred measurements. First, the volume of their weaker solution was only half that of their stronger; and, second, except for their most dilute runs, the diffusion was allowed to proceed for relatively short periods of time. It will be shown in Section F that an integral coefficient determined under these conditions differs from that obtained with the same initial concentrations, but with $V' = V''$ and $\Delta C_f/\Delta C_o = 0.5$, as in the density stirred measurements. Hartley and Runnicles reported no apparent drift in their measured D_t with time, and this is not surprising, since their more concentrated runs did not last long enough for the effect to become prominent; and, in their extended runs, the solutions were so dilute that the drift is negligible. It is possible, however, to obtain from their Table II, on the assumption that their volume ratio is 0.5, a rough but sufficiently accurate estimate of the $\Delta C_f/\Delta C_o$ in their experiments. A calculation, taking both effects into account, is then possible. The result shows that, for comparison with the density stirred data, Hartley and Runnicles' entries for the three strongest solutions in their Table III become 0.918, 0.934 and 0.950 instead of 0.922, 0.937 and 0.951, as printed. Those for their weaker solutions are unaltered. It will be noted that their value for the diffusion of 0.10 N into water has been adopted as calibration for TABLE 1. Their adjusted results are given in the last two columns.

The agreement between the experiments with density stirring, on the one hand, and those with mechanical stirring, on the other, is amazingly close; and, except for 0.02 N, is far within the precision of the measurements. It is thus evident that, for the concentrations studied here, mechanical stirring or density stirring will lead to the

same results, provided the same technique is used for the calibration and for the subsequent measurements, i.e., the stagnant layers, which were present with density stirring and absent (or practically so) with mechanical stirring, have made a reproducible contribution to the cell factor.

TABLE 1
DIFFUSION OF AQUEOUS KCl SOLUTIONS
($C_o'' = 0$; $V' = V''$; $\Delta C_f/\Delta C_o = 0.5$.)

	McB. & D.	This Research	H. & R. (corrected)	
			G-4	G-3
C_o'	D_s/D_o	D_s/D_o	D_s/D_o	D_s/D_o
0.10 N	0.918	0.918	0.918	
0.05 N	—	0.935	0.934	
0.025 N	—	0.952	0.950	
0.02 N	0.949	0.956	—	
0.01 N	0.966	0.967	0.968	0.968
0.005 N	—	—	(0.980)	0.977
0.0025 N	—	—	(0.990)	0.984

On the other hand, it is doubtful whether the same would be true under all conditions. The factors governing the thickness of the layers are not well understood, but two of them are certainly the vigor of the stirring and the viscosity. Since the stirring in the density-stirred measurements is governed primarily by the density range involved, a safe rule would seem to be that, if in the calibration and in the later measurements the density range and the viscosity are not too different, either mechanical stirring or density stirring may be used. If, however, the density range and viscosity are radically different in the calibration and the later measurements, it is probably best to eliminate the layers as completely as possible with mechanical stirring.

One obvious corollary is that, if density stirring is to be employed, care should be taken to protect the layers from random vibration. In my experience with density stirred solutions, a definite improvement in precision resulted from having the cell supports entirely independent of the walls of the air bath, thus shielding the cells from the vibration of the stirring motors.

C. THE ASSUMPTION OF A STEADY STATE IN THE DIAPHRAGM

A glance at Equation (2) shows that (strictly speaking) it cannot be correct, since, in effect, both the concentration and the concentration

gradient are specified at the surfaces of the diaphragm. Barnes[10] has given a rigorous solution of the problem for the case $V' = V'' = V$, and $C_o'' = 0$. Under these conditions, Equation (4) may be written

$$\begin{aligned} C_f' &= (C_o'/2) \ \{1 + \exp(-2DAt/lV)\} \\ C_f'' &= (C_o'/2) \cdot \{1 - \exp(-2DAt/lV)\} \end{aligned} \tag{4c}$$

He finds that Equation (4c) must be modified

$$C_f' = \frac{C_o'}{2}\left\{1 + \left(1 - \frac{\lambda^2}{180}\right) \cdot \exp\left[-\frac{2DAt}{lV}\left(1 - \frac{\lambda}{6} + \frac{\lambda^2}{45}\right)\right] + \frac{\lambda^2}{2\pi^4}\sum_{n=1}^{\infty}\frac{1}{n^4} \cdot \exp\left[\frac{-4n^2\pi^2Dt}{l^2}\right]\right\} \tag{6}$$

with an analogous expression for C_f''. Here, λ is the ratio of the volume of the solution in the diaphragm to that of the inner or outer solution. In the cells in use in the Toronto laboratory, the volume in the diaphragm is of the order, 1.5 cc, so that $\lambda < 0.02$. Thus, $\lambda^2/180$ and $\lambda^2/2\pi^4$ are entirely negligible in comparison with unity. The alteration in the form of the cell factor in Equation (6), as compared with Equation (4c), is of no consequence, since the corrected factor is also determined automatically by the calibration. It can also be verified from Barnes' equations that the distribution of concentration, under ordinary experimental conditions, differs only slightly from linearity. His calculations show that the error inherent in Equation (2) largely disappears thanks to the preliminary diffusion, and that consequently Equation (3) is valid, provided β is interpreted simply as a parameter defined by the apparatus. In brief, it is possible to treat the diaphragm cell, without sensible error, as a pseudo-steady state problem.

Barnes deals only with the case of a diffusion coefficient independent of concentration, but it would seem reasonably certain that a similar calculation (if possible) for the case of D, a function of C, would lead to a similar result. In this connection, some unpublished measurements, carried out in my laboratory by Dr. W. A. James, are pertinent. It is obvious that, for a given effective pore length, deviations from a steady state in the diaphragm will be the more serious the greater the ratio of the effective cross-sectional area of the pores to the volume of the cell. Two G-4 cells of roughly the same volume (95 cc.) and pore length (4 mm.) were selected. One had, however, only about one-half the A of the other, the cell factors being 0.1139_0 and 0.2005_5 respec-

[10] Barnes, C. Physics 5: 4. 1934.

tively. Tenth normal potassium chloride solution[11] at 25° was allowed to diffuse into an equal volume of water until the concentrations were 0.075 N and 0.025 N. This required about 4 days for the first cell and about 2 days for the second. The experiment was then repeated, this time using 0.1 N HCl instead of 0.1 N KCl; the diffusion, as before, proceeding until $\Delta C_f/\Delta C_o$ was 0.5. The ratios of the integral diffusion coefficient for an acid measurement to that for the potassium chloride measurement, immediately preceding it for three such cycles, were 1.591_0, 1.594_0 and 1.595_5 for the first cell; and 1.595_0, 1.596_0 and 1.592_5 for the second. The average ratios for the two cells, 1.593_5 and 1.594_5, are the same within the precision of the measurements, in spite of the fact that deviation from a steady state must be more serious in the second cell than in the first. Both these electrolytes show a considerable variation in diffusion coefficient with concentration in the range $0 - 0.1\ N$ (approximately 10% for KCl and 7% for HCl). Both are rapidly diffusing substances, and there is a considerable difference in their diffusion coefficients. Therefore, it seems safe to conclude that the diaphragm cell may be treated as a steady state problem, even when the diffusion coefficient is a function of concentration. The importance of this is discussed in Section F.

D. THE MECHANISM OF TRANSPORT IN THE DIAPHRAGM

Since the validity of the method rests on the assumption that the transport is diffusion controlled, it is evident that the possibility that other processes contribute is of great importance. Thus, a defective diaphragm, obviously, will permit streaming of the one solution into the other. In fact, Hartley and Runnicles' results indicate that streaming may not be entirely absent with sound G-3 diaphragm unless density differences are slight, but that it is negligible with a G-4. Various methods have been suggested for detecting imperfections in a diaphragm. Thus, Dawson[12] recommends calibrating the cell, first, in its normal position, and then tilted at an angle. Streaming, if present, will cause a change in the apparent cell factor. The best precaution, however, is always to work with a group of cells so that, if any cell has an imperfect diaphragm, this will be apparent at once in its high and erratic βD. In this connection, it should be pointed out that occasionally a cell, after months of service, will suddenly develop a leak. For this reason, if for no other, periodic recalibration is necessary.

[11] As a result of previous experiments, it was possible, by making the inner solution slightly stronger than 0.1 N, to ensure that its initial concentration at the start of the final diffusion was within a small fraction of a per cent of decinormal.

[12] Dawson, C. R. J. Am. Chem. Soc. 55: 432. 1933.

A much more disturbing possibility is that part of the transfer of solute is by surface transport along the walls of the pores. When one remembers the relatively enormous area in a typical diaphragm, it is apparent that surface transport could be a serious matter. The obvious way to settle the question is to carry out a series of measurements in the diaphragm cell on solutes whose diffusion coefficients are known from absolute measurements, and it is just here that the difficulty occurs. As will be shown in Section F, it is possible to correlate differential diffusion coefficients[13] with diaphragm cell integral data; but the vast bulk of the absolute results in the literature are integral coefficients, determined over such wide ranges of concentration that their comparison with diaphragm cell measurements is a matter of great uncertainty.[14]

This much can be said, however, of the evidence, as far as the diffusion of simple molecules and ions is concerned,—that none is available, at present, which would indicate that surface transport is sufficiently serious to invalidate the method. Thus McBain and Liu[2] showed that there is no apparent drift in the diffusion coefficient ratio for two solutes on passing from a more porous to a less porous diaphragm, or from a glass diaphragm to an alundum one. If surface transport were prominent, one would expect the ratio to be dependent on the nature and magnitude of the surface in the pores. Nevertheless, one cannot but hope that the Harned-French technique will be applied in the future to a series of solutes. One of the most interesting of these would be hydrochloric acid in dilute aqueous solution, since KCl and HCl have probably been studied more thoroughly in the diaphragm cell than any other pair of simple electrolytes.

E. CALIBRATION OF THE DIAPHRAGM CELL

Of all the questions concerning the diaphragm cell about which there has been disagreement, unquestionably the one that stands first is the choice of a standard for calibration. Mehl and Schmidt[7] made an ingenious attempt to solve the problem by filling the diaphragm with a solution of known specific conductance and determining its apparent resistance. They hoped, in this way, to fix the ratio A/l. However,

---

[13] The differential measurements of **Clack** (reference [19]) and of **Harned & French** (reference [22]) will be discussed in Section E. Other recent determinations of differential coefficients are those of **M. Randall, B. Longtin, & H. Weber** (J. Phys Chem. **45**: 343. 1941) and of **W. G. Eversole, H. M. Kindsvater, & J. D. Petersen** (J Phys. Chem. **46**: 370. 1942).

[14] The measurements of **W. A. Patterson & J. T. Burt-Gerrans** (Can. J. Research **22**: 5. 1944) are an interesting exception. They studied the diffusion of copper sulphate in the presence of a very large excess of sulphuric acid. Under these conditions, the diffusion coefficient of the copper sulphate is independent of concentration. Their results agree within the precision of their measurements with the diaphragm cell data of **Cole & Gordon** (reference [20]).

the difficulty of obtaining an accurate value for the resistance under these conditions is great, and it is unlikely that an A/l, measured in this way, is valid for diffusion measurements in which stagnant layers may be present. Most experimenters, therefore, have followed Northrop and Anson's example [1] and determined the cell factor by calibration with a solute of known diffusion coefficient. In their original paper, Northrop and Anson state that they used for this purpose hydrochloric acid, lactose and several salts. The example they give uses a value for the diffusion of 0.1 N hydrochloric acid into water obtained by extrapolation from Öholm's data. Later,[15] recognizing, the desirability of calibrating with a solute whose diffusion coefficient was independent of concentration, they recommended 2 N NaCl. While it is true that D for NaCl changes much less at moderate concentrations than is the case for most electrolytes, it is nevertheless far from constant,[16] and, therefore, it is a little difficult to see the advantage of such a standard.

McBain and his associates [3] based their measurements on Cohen and Bruins' value[17] of D for the diffusion of 0.1 N KCl into water at 20°, a procedure that has been widely followed (by myself, among others, in some of our earlier work). In 1937,[18] I pointed out that Cohen and Bruins' value was an integral coefficient, which, owing to the design of their apparatus, was essentially of a different nature from that obtained in a diaphragm cell measurement for the same initial concentrations. Subsequently, Hartley and Runnicles [4] considered the question in detail, and showed that Cohen and Bruins' result corresponded more nearly to the integral coefficient obtained with a diaphragm cell when 0.06 N KCl diffused into water. They also pointed out that Cohen and Bruins' value was almost certainly too high, owing to the inevitable mixing which must occur in their technique when the diffusion column is separated into layers at the end of the experiment.

There have been two attempts to select a standard on *a priori* grounds, the first by myself,[18] and the second by Hartley and Runnicles.[4] The two are alike in that both assume the limiting Nernst value for D_0, the diffusion coefficient at infinite dilution, and both recognize the distinction between differential and integral coefficients. They differ in the method of extrapolation. The former is based on a semi-empirical relation which represented Clack's differential data [19]

---

[15] Anson, M. L., & J. H. Northrop. J. Gen. Physiol 20: 575. 1937.

[16] Between 0.4 N, at which D is a minimum, and 2 N, D changes by about 5%, while the value at 0.4 N is some 10% less than the Nernst value at infinite dilution (See reference [20]).

[17] Cohen, E., & H. R. Bruins. Zeit. f. phys. Chem. 103: 337. 1923.

[18] Gordon, A. R. J. Chem. Phys. 5: 522. 1937.

[19] Clack, B. W. Proc. Phys. Soc. 36: 313. 1924.

for KCl, NaCl and KNO_3 moderately well. The latter postulates that D/D_0 must approach asymptotically the limiting equation of Onsager and Fuoss.[20] The two methods yield not very different results, which is not surprising, since both assume that the principal reason for the variation of the differential coefficient with concentration in dilute KCl solution lies in the thermodynamic factor $(1 + d \ln f/d \ln C)$. My proposed standard was $D_i = 1.836 \times 10^{-5}$ cm²/sec. when 0.1 N KCl diffuses at 25° into an equal volume of water, and the diffusion is allowed to proceed until the concentrations of the inner and outer solutions are 0.075 N and 0.025 N respectively, i.e., $\Delta C_f/\Delta C_0$ was 0.5. This was subsequently revised [21] for the same experimental conditions to 1.842×10^{-5}. Hartley and Runnicles recommend for experiments of very short duration, for which $\Delta C_f/\Delta C_0$ is only slightly less than unity, $D_i/D_0 = 0.922$ for the diffusion of 0.1 N KCl into water. It will be shown in Section F that this corresponds, for experiments in which $\Delta C_f/\Delta C_0 = 0.5$ and $V' = V''$, to $D_i/D_0 = 0.918$, which is the value used for calibration in TABLE 1. If this be combined with the Nernst D_0 for 25°, viz., 1.994×10^{-5} cm²/sec., there results for the conditions of my proposed standard 1.830×10^{-5}, a value differing by 0.6% from that on which the Toronto measurements were based.

It must be admitted that the practical application of the Hartley and Runnicles standard rests on the validity of the Nernst limiting equation, and it is probably not generally recognized how extremely subjective was the experimental evidence in favor of the equation until the measurements of Harned and French [22] were reported. The evidence usually advanced is obtained from Öholm's data,[23] which have been quoted in a review article by Williams and Cady.[24] Öholm diffused 0.01 N solutions into water, and found that the resulting integral coefficients were in reasonable agreement with the limiting Nernst values, in the case of potassium, sodium and lithium chlorides, and potassium iodide. With hydrochloric acid and potassium and sodium hydroxides, however, his results were from 6% to 10% less than the limiting coefficients. Diaphragm cell results [21] in my laboratory also showed, that if the Nernst equation were valid for KCl, it could not (on the basis of any reasonable extrapolation) be valid for HCl. Another disquieting feature of Öholm's data is that his integral values presumably correspond, owing to the design of his apparatus, to the differential coefficients for concentrations in the neighborhood of

[20] **Onsager, L., & R. M. Fuoss.** J. Phys. Chem. **36:** 2689. 1932
[21] **James, W. A., E. A. Hollingshead, & A. R. Gordon.** J Chem Phys **7:** 89. 1939.
[22] **Harned, H. S., & D. M. French.** Ann. N. Y. Acad Sci. **46** (5) 280 1945.
[23] **Öholm, L.** Zeit. f. phys. Chem. **50:** 309. 1905.
[24] **Williams, J. W., & L. C. Cady.** Chem. Rev. **14:** 171. 1934.

0.003 N. All evidence, both theoretical[20] and experimental,[22] indicates that for strong 1-1 electrolytes, in very dilute solution, the differential coefficient is less than the limiting value, so that Öholm's data for the salts, to be consistent with the limiting coefficients, should be, on the average, two or three per cent [20] less than the limiting values. Of course, this can be "explained" on the basis that Öholm's technique tended to give slightly high results owing to the method of sampling he used, but the same argument would apply equally to the acid and the hydroxides, and would still leave a divergence in their cases from the limiting equation. Nevertheless, I believe that the evidence of Harned and French indicates the validity of the Nernst equation for potassium chloride, and, consequently, justifies its use in arriving at an absolute value from Hartley and Runnicles' extrapolation.

If Hartley and Runnicles' calibration be adopted, it will be shown that the values of D_i/D_o, entered in TABLE 1, correspond for the differential coefficient D to the relation

$$D/D_o = 1 - 0.515\sqrt{C} + 0.59C \tag{7}$$

Equation (7) is of course entirely empirical; but, as will be shown in Section F, is consistent with the integral data of the table, and is valid over a moderate range of temperature. Clack's results [19] are unambiguous differential coefficients which may be compared with Equation (7). For his most dilute solution, *viz.*, 0.08 N at 18.5°, he gives $10^5 \cdot D = 1.526$ and 1.539. The mean after correction to a diffusion relative to solution basis, is $10^5 \cdot D = 1.529 \pm 0.007$. Since the Nernst[25] D_o, at 18.5°, is 1.703×10^{-5}, Equation (7) predicts for 0.08 N, at this temperature, 1.535×10^{-5}, which differs from Clack's result by no more than the apparent precision of his measurements.

The strongest evidence as to the correct standard, however, is provided by the results of Harned and French.[22] At this point, I particularly wish to express my thanks to Professor Harned for making these data available to me prior to publication. It should be emphasized that Harned and French's results, like Clack's, are true differential coefficients. For 25°, they find:

$\sqrt{C}$	0.05000	0.05922	0.05899	0.07050
$10^5 \cdot D$, H. & F.	1.944	1.953	1.921	1.913
$10^5 \cdot D$, Equation (7)	1.945	1.937	1.937	1.927

The values[*] computed from Equation (7) and the Nernst D_o for 25°, *viz.*, 1.994×10^{-5}, are tabulated for comparison. It is at once ap-

[*] This was computed from the limiting mobilities obtained by interpolation in Gunning & Gordon's Tables V and VI (J. Chem. Phys. 12: 126. 1942).

parent that Equation (7) is consistent with Harned and French's measurements, within their apparent precision.

The position with respect to standards is then briefly this: the ratios of TABLE 1 are the result of three entirely independent researches with the diaphragm cell, and are arbitrary only to the extent of the calibration, common to all the entries in the table. It has been shown here that Hartley and Runnicles' calibration is consistent with Clack's results, on the one hand, and with those of Harned and French, on the other. While, admittedly, the precision of the Harned and French technique will be improved in the future, it is difficult to see where there can be any serious systematic error in the method, and, consequently, their results, taken in conjunction with Clack's, indicate that Hartley and Runnicles' standard cannot be seriously in error. It would, therefore, seem best to use, as a standard for calibration ,0.1 N KCl diffusing into water at 25° with $D_t = 1.838 \times 10^{-5}$ cm²/sec. when $\Delta C_f/\Delta C_o$ is nearly unity, or $D_t = 1.830 \times 10^{-5}$ when $\Delta C_f/\Delta C_o = 0.5$, and equal volumes of solution and water are used. In any event, the use of Cohen and Bruins' value is entirely without justification.

It should be noted that, to be consistent with this calibration, the data previously reported from my laboratory for hydrochloric and sulphuric acids[21, 26, 27] and calcium chloride[28] must be decreased by 0.6%, while the results for copper sulphate,[29] which were based on Cohen and Bruins' calibration, must be decreased by 3%.

F. THE RELATION BETWEEN THE DIAPHRAGM CELL INTEGRAL COEFFICIENT AND THE DIFFERENTIAL COEFFICIENT

One of the most curious phenomena in the whole field of diffusion has been the failure of many experimenters to recognize the distinction between the differential diffusion coefficient, valid for a concentration, and the integral coefficient which results from a measurement covering a range of concentrations. The resulting confusion has had some unfortunate consequences, since the "diffusion constant" is usually only a constant in a Pickwickian sense of the term. Thus, one finds the quantity determined in an experiment in which a solution, initially decinormal, is allowed to diffuse into water, referred to as "the diffusion constant for 0.1 N". The fact that it could, with equal justice, be spoken of as the value at infinite dilution is conveniently ignored.

[26] James, W. A., & A. R. Gordon. J. Chem. Phys. 7: 963. 1939.
[27] Hollingshead, E. A., & A. R. Gordon. J. Chem. Phys. 8: 423. 1940.
[28] Hollingshead, E. A., & A. R. Gordon. J. Chem. Phys. 9: 152. 1941.
[29] Cole, A. F. W., & A. R. Gordon. J. Phys. Chem. 40: 733. 1936.

If, as is usually done, the diffusion is defined in terms of transport relative to a fixed column of solution,[30] the differential coefficient D for a concentration C is given in the case of unidimensional diffusion in the x direction by

$$dq/dt = -D \cdot \partial C/\partial x \tag{8}$$

where dq/dt is the amount of solute diffusing per second through a horizontal plane, one square centimeter in area, at a height x in the column of solution, the concentration being C in the plane at x. Alternatively, Equation (8) may be written

$$\frac{\partial C}{\partial t} = \frac{\partial}{\partial x}\left(D \cdot \frac{\partial C}{\partial x}\right) \tag{9}$$

It is only when D is not a function of C, and, therefore, is independent of x, that Equation (9) reduces to the usual Fick equation

$$\partial C/\partial t = D \cdot \partial^2 C/\partial x^2 \tag{10}$$

In contrast to Equation (9), Equation (10) can be readily integrated with suitable boundary conditions, and the resulting forms have been widely used, even when D is not independent of C, to obtain from experimental data values of the diffusion coefficient. It is evident, however, that the resulting quantity is an averaged or integral coefficient which will depend, not only on the total range of concentration involved, but also on the duration of the experiment, and, through the boundary conditions, on the type of apparatus. This much can be said, however,—an integral coefficient, computed by means of the appropriate solution of Equation (10), will, in general, be equal to the differential coefficient for some concentration lying in the range used. It is in this connection that one of the great advantages of the diaphragm cell method becomes apparent, since it is possible for a diaphragm cell measurement to determine just what this concentration is. In what follows, it will be assumed that the differential diffusion coefficient D is a function of concentration; in particular, that

$$D/D_0 = 1 + f(C) \tag{11}$$

where D_0 is the coefficient at infinite dilution.

It is at once apparent that if Equation (11) be true, the derivation of Equation (4) is invalid. Nevertheless, it is convenient to define the diaphragm cell integral coefficient D_t by means of Equation (4), since it can always be computed from the duration of the experiment and the known values of the cell factor and of $\Delta C_f/\Delta C_0$. A more general

[30] In this connection, see references [4, 19] and [29].

theory, however, may be developed without difficulty, since (as was shown in Section C) the diaphragm cell may be treated as a steady state problem. For a steady state, the quantity which is constant for a particular instant of the diffusion from the top of the diaphragm ($x = 0$) to the bottom ($x = l$) is $D \cdot \partial C/\partial x$, since, otherwise, there would not be the same transport through every cross section from top to bottom. If, at this time, the inner and outer concentrations are C' and C'', it is then possible to define, for this instant, an effective diffusion coefficient D' in terms of the flow:

$$-D'(C' - C'')/l = D \cdot \partial C/\partial x \tag{12}$$

Multiplication by dx and integration from $x = 0$ to $x = l$ gives

$$D'/D_o = 1 + (1/\Delta C) \int_{C''}^{C'} f(C) \cdot dC \tag{13}$$

where, as before, ΔC stands for $(C' - C'')$. A comparison of Equation (12) with Equation (1) and Equation (2) shows that Equation (3) must now be modified to give

$$d(\Delta C)/(\Delta C) + \beta D' dt = 0 \tag{14}$$

where, however, D' will now change as the diffusion proceeds, owing to its dependence on concentration To proceed to the integrated form of Equation (14), define a quantity $F(C', C'')$ by

$$1 + F(C', C'') = D_o/D' \tag{15}$$

Substitute for D' in Equation (14) by means of Equation (15), rearrange and integrate. There results

$$\ln \frac{\Delta C_f}{\Delta C_o} + \int_{\Delta C_o}^{\Delta C_f} \frac{F(C', C'')}{\Delta C} \cdot d(\Delta C) = -\beta D_o t \tag{16}$$

But the leading term on the left of Equation (16) is $-\beta D_i t$; hence,

$$D_i/D_o = 1 + (1/\beta D_o t) \int_{\Delta C_o}^{\Delta C_f} \frac{F(C', C'')}{\Delta C} \cdot d(\Delta C) \tag{17}$$

Moreover, if C_i be defined as the concentration for which the integral coefficient D_i is equal to the differential coefficient D, it follows that $f(C_i)$ must be identically equal to the second term on the right of Equation (17), thus making it possible to solve for C_i. An equation equivalent to Equation (17) was deduced in a paper of mine [18] appearing in 1937, but the form was not nearly so convenient to handle in practice as Equation (17).

If the differential coefficient be known as a function of concentration, Equation (17) serves to predict the integral coefficient for a diaphragm

cell measurement. As far as the actual mechanics of the calculation are concerned, the integration of Equation (13) for given values of C' and C'' can be performed at once, analytically, graphically, or tabularly. For example, in an experiment in which 0.1 N solution diffuses into an equal volume of water, the integral might be computed for ($C' = 0.100$, $C'' = 0$), ($C' = 0.098$, $C'' = 0.002$) and so on. If the volume of the water were twice that of the solution, the values selected might be ($C' = 0.100$, $C'' = 0$), ($C' = 0.098$, $C'' = 0.001$) etc. In any event, from each value of the integral in Equation (13), the corresponding value of F (C', C'') and of the integrand in Equation (17) would be calculated, and the integration in Equation (17) then effected graphically or tabularly.

There is one special case of Equation (17), which is of some importance. If the differential coefficient be linear in the concentration, i.e., if $f(C) = BC$, Equation (13) becomes

$$D'/D_o = 1 + B(C' + C'')/2 \tag{18}$$

If equal volumes [31] of inner and outer solution be used, ($C' + C''$), and, consequently, D', will be constant throughout the experiment; hence,

$$D_t = D' = D(\bar{C}) \tag{19}$$

where $D(\bar{C})$ stands for the differential coefficient for the mean concentration of the experiment $\bar{C} = (C_o' + C_o'')/2 = (C_f' + C_f'')/2$. Equation (19) was first deduced by Cole and Gordon, [29] and is a moderate approximation, even in cases where there is a non-linear dependence of D on C, provided the concentration range is not too great, and provided D is monotonic in that range.

If Equation (17) is a solution of the problem, "Given D, find D_t", the inverse, and much more usual problem, "Given D_t, find D", involves a short series of approximations. The method can be illustrated by showing how the differential coefficients may be obtained from the integral data of TABLE 1. The second column of TABLE 2 gives the integral coefficients obtained from the ratios of TABLE 1 and the Nernst D_o for 25°, 1.994 × 10^{-5} cm²/sec. In computing the entries, the ratios used were averages for a given C_o', except that McBain and Dawson's value for 0.02 N and the bracketed entries were omitted. The first approximation consists in identifying the integral coefficients with the differential coefficients for the mean concentrations of the experiments, entered in the third column. Note that this approxima-

[31] See Section A.

tion is only possible owing to the condition that equal volumes were used for the inner and outer solutions. See Equation (18). On the basis of this identification, it is possible to represent the data by means of the empirical equation

$$D/D_0 = 1 - 0.515\sqrt{C} + 0.66C \qquad (20)$$

as is shown by the entries in the column headed "$10^5 \cdot D$, Calc. 20." The next step is to compute the D_t for the various experiments by means of Equations (13), (17), and (20). The results of this calculation are entered in the fifth column and the first entry may be taken as an ex-

Table 2

Diffusion of Aqueous KCL Solutions

$C_0'' = 0$, $V' = V''$; $\Delta C_f/\Delta C_0 = 0.5$, 25° C

		First Approximation			Second Approximation		
C_0'	$10^5\ D_t$ (obs.)	C_t	$10^5\ D$ (Calc 20)	$10^5\ D_t$ (Calc 20)	C_t	$10^5\ D$ (Calc. 21)	$10^5\ D_t$ (Calc. 21)
0.10	1 830	0.05	1.830	1.836	0.04423	1.830	1.829
0.05	1.863	0.025	1.864	1.869	0.02284	1.865	1.865
0.025	1.896	0.0125	1.896	1.899	0.01159	1.896	1.896
0.02	1 906	0.01	1.904	1.907	0.00930	1.906	1.906
0.01	1.928	0.005	1.928	1 930	0.00468	1.929	1.929
0.005	1.948	0.0025	1.946	1 947	0.00235	1.947	1.947
0 0025	1 962	0 00125	1.959	1.960	0.00118	1.960	1.960

ample. The leading term on the left of Equation (16) is $-\ln 2 = -0.69315$. The second term, computed by tabular integration, is -0.05955. Hence, $\beta D_0 t = 0.75270$, and D_t/D_0 is 0.92088. (In this example, the computation has been carried to an excessive number of significant figures for purposes of comparison with an approximation discussed below). It will be noted that there are slight discrepancies between the corresponding entries in the fourth and fifth columns, which would, however, only be significant in work of the highest precision.

The next step is to substitute the calculated D_t of the fifth column in the quadratic (20) and solve for C_t, the concentrations for which the integral coefficients are identical with the differential (on the assumption that Equation (20) is valid). The resulting concentrations are given in the sixth column, and the second approximation consists in identifying the observed integral values of the second column with the

differential values for these new concentrations. On this basis, the differential coefficients can be represented by

$$D/D_o = 1 - 0.515\sqrt{C} + 0.59C \tag{21}$$

It will be noted that Equation (21) is identical with Equation (7), which was employed in the previous section in the discussion of the cell factor. The column headed "$10^5 \cdot D$ Calc. 21" shows, on comparison with the second column, that Equation (21) adequately represents the data. The final step in this approximation, and a preliminary for a third approximation, if necessary, is to compute the D_i as before, but this time to use Equation (21) for the differential coefficients instead of Equation (20). The results are given in the last column of the table, and it is at once evident that a third approximation is unnecessary. In brief, it has been possible to arrive at an expression for the differential coefficients, *viz.*, Equation (21), which is consistent with the observed integral data. It should also be noted that the only simplifying assumption involved is the entirely justifiable one that the diaphragm cell can be treated as a steady state problem.

While the procedure outlined above can always be applied, it is, at best, laborious, and the question naturally arises whether a simpler but adequate approximation is available. A hint is given by the fact that $F(C', C'')$ does not change greatly as the diffusion proceeds. For example, in the sample calculation discussed above for 0.1 N, $F(C', C'')$, computed from Equation (20), is 0.08175, when $C' = 0.100$, $C'' = 0$; and is 0.08802, when $C' = 0.075$, $C'' = 0.025$. This suggests that a constant value F_m be used throughout the integration of Equations (16) and (17) where F_m is defined by

$$\frac{1}{1 + F_m} = 1 + (1/\Delta C_m) \cdot \int_{C_m''}^{C_m'} f(C) \cdot dC \tag{22}$$

and $C_m' = (C_o' + C_f')/2$, $C_m'' = (C_o'' + C_f'')/2$. In this example, C_m' would be 0.0875 N and C_m'' 0.0125 N. The second item on the left of Equation (16) can then be integrated at once, and the equation rearranged to give, instead of Equation (17),

$$D_i/D_o = 1 + (1/\Delta C_m) \cdot \int_{C_m''}^{C_m'} f(C) \cdot dC \tag{23}$$

Equation (23), gives in the case chosen for illustration, $D_i/D_o = 0.92091$, which may be compared with the result previously obtained from Equation (17), $D_i/D_o = 0.92088$. A number of tests of Equation (23) show that the error introduced by its use in place of Equation (17) is at most 0.02% in experiments where there is a variation of not more than 20%

in $F(C', C'')$ during the course of the diffusion. The immense saving in labor resulting from the use of Equation (22) is obvious.

Before leaving this question, it is necessary to consider the effect of the experimental conditions on the diaphragm cell integral coefficient, and this is illustrated in TABLE 3, which shows the effect of the volume ratio and of the duration of the experiment on the integral coefficient. In computing the entries, Equations (21) and (23) were employed. Calculations similar to these served to convert Hartley and Runnicles' data to the conditions of the density stirred measurements summarized in TABLE 1. For example, their conditions for the experiment involving 0.1 N solution corresponds to the first entry in the second line in TABLE 3, while the density stirred measurements correspond to the last entry in the top line. The table shows that a mere statement of the initial concentrations in a diaphragm cell measurement is not necessarily sufficient to define the experimental conditions. However, it should be pointed out that the case chosen for illustration is rather an extreme one. For example, a similar table for the diffusion of 0.005 N into water would have 0.9772 and 0.9764 as first and last entries in the top line, and 0.9772 and 0.9754 as corresponding entries in the second line. Even so, it would be helpful in correlating data if experimenters always stated explicitly the volume ratio and the final concentrations in their experiments, even if they did not wish to carry out calculations of the type discussed in this section.

TABLE 3
D_s/D_o FOR AQUEOUS KCL SOLUTIONS
($C_o' = 0.1\ N$; $C_o'' = 0$; 25° C)

$\Delta C_f/\Delta C_o$	1.0	0.9	0.8	0.7	0.6	0.5
$V' = V''$	0.9209	0.9199	0.9191	0.9185	0.9179	0.9174
$V' = 2V''$	0.9209	0.9192	0.9179	0.9167	0.9157	0.9147
$V' = V''/3$	0.9209	0.9210	0.9212	0.9214	0.9217	0.9221

It is sixteen years since Northrop and Anson's paper appeared, and it is now possible to make an assessment of the state of the art for the diaphragm cell. It is still unsurpassed in its simplicity and in the precision of the data it yields. The only question that can be raised concerns the accuracy of the results, and, in particular, the systematic errors which may be inherent in the method. Of these, the most serious would be transport through the diaphragm by processes other than diffusion. The evidence is of a somewhat negative nature, and before this question can be considered closed, additional absolute differential

data are needed to serve as a check on diaphragm cell results. The choice of the correct absolute standard for calibration is still open, and, for the moment, that recommended in Section E, is in my opinion, the best that can be made. Here, again, further absolute data would be helpful. Finally, I believe the discussion in Section F places the theory of the method, when the diffusion coefficient is a function of concentration, on a satisfactory basis.

DIFFUSION CONSTANT MEASUREMENT IN THEORY AND PRACTICE

By Edward M. Bevilacqua, Ellen B. Bevilacqua, Margaret M. Bender and J. W. Williams

From the Laboratory of Physical Chemistry
University of Wisconsin, Madison

INTRODUCTION

The pioneering work of Svedberg in developing the ultracentrifuge has led to the accompanying development of methods for measuring the diffusion constants of macromolecular substances in solution. The popularity of the sedimentation velocity ultracentrifuge is due to the fact that it permits rapid determination of the sedimentation rates of these substances in high centrifugal fields. It is well known, however, that the sedimentation rate is not a thermodynamic property and cannot be interpreted in terms of molecular weight without additional information. In principle, measurements of either viscosity or of free diffusion constant may be used to supply this additional information To date, the latter has been used almost exclusively and there are in the literature extensive tables of data on proteins and some of the other macromolecular materials from which molecular weights have been calculated by the use of the familiar equation

$$M = \frac{RTs}{D(1 - \bar{v}\rho)}$$

Unfortunately, some of these data have required substantial revision within a relatively short period, and we believe that this is largely due to uncertainties in the magnitudes of the diffusion constants used, even after allowance is made for the fact that, in early work, these were estimated from the spreading of the sedimenting boundary. Indeed, when the diffusion constant data of the literature are examined it becomes apparent that some repetition of the work is required, especially that which is based on single experiments. It is generally recognized that, for many materials which have highly asymmetrical molecules in solution, the present experimental methods are not sufficiently sensitive to be used at dilutions high enough to ensure ideal behavior, but they are usually considered adequate for materials whose molecules are sym-

metrical in solution. Thus, we find that, in several recent reviews, it is recognized that published values of the diffusion constants of asymmetrical materials must be considered as first approximations, but the estimates of the accuracy of results from the study of the diffusion of such materials as the globular proteins are optimistic.

In this report, besides describing some of the experimental results obtained in this Laboratory, we wish to recall attention to some of the anomalies observed in diffusion experiments and to emphasize the necessity for adequate estimation of the precision of the data reported.

PART I. THEORY

Most of the recent experiments on free diffusion in solution, particularly those made for characterizing proteins, have made use of one of the refractometric methods for following the progress of the diffusion. The development of the scale method of Lamm[1] and the rediscovery and improvement of the schlieren method used by Wiener[2] and by Thovert[3,4] led to experimental results which demonstrated the superiority, in both convenience and accuracy, of these methods over those which require interruption of the diffusion for analysis or in which concentrations are estimated from measurements of light absorption.

In both of these methods, the ordinates of the experimental curves obtained are proportional to the index of refraction gradient at each point in the cell. To calculate diffusion constants from these, use is made of the solution of Fick's second Law

$$\frac{\partial c}{\partial t} = \frac{D\partial^2 c}{\partial x^2} \tag{1}$$

where c = concentration

t = time

x = distance in the direction of diffusion

For diffusion from a dilute solution into pure solvent, this may be written[5]

$$\frac{dn}{dx} = \frac{n_1 - n_0}{2\sqrt{\pi D t}} e^{\frac{-x^2}{4Dt}} \tag{2}$$

[1] Lamm, O. Nova Acta Soc. Sci. Upsala 4 (10): 6. 1937.
[2] Wiener, O. Wied. Ann. 49: 105. 1893.
[3] Thovert, J. Ann. Phys. (9) 2: 369. 1914.
[4] These methods have been reviewed by Bridgman, W. B., & J. W. Williams. Ann. N. Y. Acad. Sci. [illegible] (5): 195. 1942.
[5] Williams, J. W., & L. C. Cady. Chem. Rev. 14: 171. 1934.

where $\frac{dn}{dx}$ = index of refraction gradient

n_1 = index of refraction of solution

n_0 = index of refraction of solvent

D = diffusion constant

This expression was derived by Wiener, who assumed that the index of refraction of the solution is a linear function of the concentration of the solute. It has been shown[1] that the derivation is proper for dilute solutions even without this assumption.

In the scale method, the experimental curve is a plot of the displacement of the lines of a uniform scale photographed through the system in which diffusion is taking place as a function of the distance of these lines from an arbitrary reference line. For the calculation, the best curve is drawn through these points and traced onto another sheet with an arbitrary origin chosen for convenience. Scales of abscissae and ordinates are chosen and a table of ordinates (S) for equal intervals on the axis of abscissae (s) constructed. Since the origin is located by estimating the centroid in tracing the curve, a generalized form of Equation (2) is used, which takes into account the distance of the true centroid from the origin

$$S = \frac{N\omega}{\sigma\sqrt{2\pi}} e^{-\frac{(s-\beta)^2\omega^2}{2\sigma^2}} \tag{3}$$

where S = scale line displacement corresponding to scale line distance s from chosen origin.

$\sigma^2 = 2\,Dt$

β = distance of chosen origin from centroid

ω^2 = numerical constant

$N\omega$ = area of curve

If this analytical expression is the equation for the experimental curve, values of D may be calculated from the s,S table, using the properties of the curve. The two most frequently used equations for D are based on the height and area (4a) and the second moment and area of the curve (4b)

$$D = \frac{(\Sigma S)^2}{S^2_{max.}} \frac{\omega^2}{2\pi t} \tag{4a}$$

$$D = \left[\frac{\Sigma s^2 S}{\Sigma S} - \left(\frac{\Sigma s S}{\Sigma S}\right)^2\right] \frac{\omega^2}{2t} \tag{4b}$$

Since the units of s and S in Equation (3) are arbitrary, it is necessary to convert the values from different experiments to the same (dimensionless) units in order to be able to compare them. This is usually done by the transformations

$$\begin{aligned} \Xi &= \left[\frac{5\sigma}{\omega \Sigma S}\right] S \\ \xi &= \frac{(s - \beta)\omega}{2\sigma} \end{aligned} \tag{5}$$

The resulting set of values of Ξ and ξ may then be compared with those from other experiments and with the corresponding normal (Gauss) curve

$$\Xi = \frac{5}{\sqrt{2\pi}} e^{-2\xi^2} \tag{6}$$

Coincidence of the experimental points with the normal curve shows that Equation (2) applies to the diffusion system studied.

The assumption that Equation (2) does describe the variation of index of refraction gradient (and so of concentration gradient) as a function of distance in the diffusion cell and of time, so long as diffusion has not progressed far enough to affect concentrations at the end of the cell, has been justified by a great mass of experimental data. As a result, coincidence of the experimental with the normal curve is taken as an indication of homogeneity, or, more strictly, no inference about the solute is drawn from coincidence, but deviations of the experimental curve from the normal are attributed to some abnormality of the solute. Properties of the solute producing such deviations will, of course, produce deviations from the expected behavior, when methods based on the analysis of the solution are used to follow a diffusion experiment, and the earlier attempts to find quantitative explanations for the observed phenomena were made by workers using these methods.[6, 7] More recently, attempts have been made to make quantitative calculations from the results of experiments using refractometric methods.

In briefly reviewing the most frequently observed deviations from the Gauss normal curve, we shall divide them into the obvious classes: (A) symmetrical curves which are not normal; (B) asymmetrical curves; and (C) other abnormalities.

[6] Krüger, D., & H. Grunsky. Z. physik Chem. A170: 161. 1934.
[7] Herzog, R. O., & H. Kudar. Z. physik Chem A167: 343. 1933

(A) Symmetrical Curves which are not Normal

It may easily be shown that the sum of two or more normal curves is not normal: Rewriting Equation (2)

$$y_i = \frac{c_i}{\sigma_i\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma_i^2}} \qquad y_i = \frac{\partial n}{\partial x}$$

$$c_i = n_1 - n_0$$

$$\sigma_i = \sqrt{2D_i t} \tag{7}$$

The sum of a set of y_i is

$$y = \Sigma y_i = \frac{1}{\sqrt{2\pi}} \sum \frac{c_i}{\sigma_i} e^{\frac{-x^2}{2\sigma_i^2}} \tag{8}$$

For this to be a normal curve, there must be a c_0 and a σ_0 so that

$$y = \Sigma y_i = \frac{c_0}{\sqrt{2\pi}\sigma_0} e^{-\frac{x^2}{2\sigma_0^2}} \tag{9}$$

We may set all c_i's equal, so that $c_0 = nc_i$. Expanding the right members of Equations (8) and (9) in series and equating the coefficients of like powers of x we obtain the desired conditions

$$\begin{aligned} \frac{n}{\sigma_0} &= \sum \frac{1}{\sigma_i} \\ \frac{n}{\sigma_0^3} &= \sum \frac{1}{\sigma_i^3} \\ \frac{n}{\sigma_0^5} &= \sum \frac{1}{\sigma_i^5} \quad \text{etc.} \end{aligned} \tag{10}$$

which is only possible for the trivial case $\sigma_0 = \sigma_1 = \sigma_2 = \;.\;.\; \sigma_m$.

It follows immediately that a possible explanation for the occurrence of symmetrical non-Gaussian curves is the presence of a mixture, each component of which diffuses independently. Pearson[8] showed that such curves may be analyzed by making use of the moments of the experimental curve

$$\nu_i = \int_{-\alpha}^{\alpha} x^i y dx \tag{11}$$

For the simplest case, a curve which is the sum of two normal curves, a set of equations involving the even moments up to the sixth gives the two standard deviations and the concentrations. Gralén, in applying Pearson's method to the particular problem of diffusion, considered only one half of the curve, since it is symmetrical. He demonstrated that,

[8] Pearson, K. Phil. Trans. Roy. Soc. A185: 71. 1894.

although any pair of moments of a normal curve may be used to calculate the standard deviation (and so D), successively higher values of D are obtained as higher pairs of moments of a non-normal curve are used Since the calculated values are weighted averages, analogous to the various average molecular weights, Gralén[9] suggested that the ratio of the value of D, calculated by the second moment method (4b) to that calculated from the height and area (4a) of the curve, be used as a criterion of homogeneity. This parallels the characteristic constant suggested by Lansing and Kraemer,[10] which is the ratio of weight average to number average molecular weight. The deviation of both constants from one is a measure of the distribution of sizes in a substance.

Neurath[11] has suggested a more specific application of Pearson's original method, proposing to analyze the experimental curve arising from the diffusion of a mixture believed to consist of only two substances; in particular, a protein containing an impurity. Simplifying the calculation by using only the positive half of the curve, he obtained an expression for the diffusion constants of the two substances involving only the area and the first, second, and third moments.

Since any of these procedures for analyzing experimental curves would be extremely useful if it could be practically applied in diffusion studies, attempts were made to use them on the curves obtained from the diffusion of known mixtures. Of the results of Pearson, Gralén, and Neurath, those of the latter seemed of most interest, and were considered first. Writing Equation (8) for two components, there was obtained the expression:

$$\sigma_{1,2} = -\sqrt{\frac{\pi}{8}}\,\frac{2\nu_1\nu_2 - \nu_0\nu_3}{2\nu_0\nu_2 - \pi\nu_1^2} \pm \sqrt{\frac{\pi}{8}\left(\frac{2\nu_1\nu_2 - \nu_0\nu_3}{2\nu_0\nu_2 - \pi\nu_1^2}\right)^2 - \frac{\pi\nu_1\nu_3 - 4\nu_2^2}{4\nu_0\nu_2 - 2\pi\nu_1^2}} \qquad (12)$$

The first calculations using this equation were made on the curves obtained by allowing mixtures of KCl and sucrose to diffuse. For a KCl to sucrose ratio of 1:10, the experimental curves could not be distinguished from the normal curve, and, at higher ratios, the analysis gave results which were considerably in error. It was suggested that the choice of KCl as one component might be objectionable, because a very slight deviation from the normal is observed when it diffuses alone. A synthetic curve was, therefore, constructed and the analysis made in the usual manner. In this case, the values of σ_1 and σ_2 were within 15 and 25 per cent, respectively, of the correct values, but even so good

[9a] Gralén, N. Kolloid Z 95: 188. 1941
[9b] Gralén, N. Dissertation. Upsala 1944.
[10] Lansing, W. D. & E. O. Kraemer. J. Am. Chem. Soc. 57: 1369. 1935.
[11] Neurath, H. Chem. Rev. 30: 357. 1942.

a correspondence as this represents a surprisingly large deviation, since the error in the curve analyzed amounts essentially to the width of the pencil line used to draw it

In view of this result, an attempt was made to calculate the expected error* in the values of σ_1 and σ_2 corresponding to a given error in the individual values of y. See Equation (8). The major difficulty in this procedure is in assigning the source of error. The experimental curve obtained (in the scale method) is simply a series of points, presumably randomly scattered about the true curve. From this set of points, both the curve and the axis of abscissae (base line) are located. We have, therefore, a possibly constant error due to the base line as well as the random error in the points of the curve. Since, however, this source of error is usually taken into account in routine work and adjusted by comparison of the areas of the whole set of curves for an experiment, we shall not consider it here.

The method for calculation of the error in any property of a frequency curve was developed by Pearson,[12] who assumed a binomial distribution of deviations from the correct frequency, when a sample is taken from a given universe. This is probably considerably broader than the distribution occurring in diffusion curves. It is believed that the error in any value of y is independent of y (and of x) and we have used this assumption in following Pearson's procedure. We have from Equation (12)

$$\sigma_{1,2} = f(\nu_0, \nu_1, \nu_2, \nu_3). \tag{13}$$

For small fluctuations

$$d\sigma_{1,2} = f_0 d\nu_0 + f_1 d\nu_1 + f_2 d\nu_2 + f_3 d\nu_3 \tag{14}$$

where f_i is the partial derivative of $f(\nu_0, \nu_1, \nu_2, \nu_3)$ with respect to the moment indicated by the subscript. Squaring, summing for all fluctuations, and dividing by the total number of fluctuations

$$\begin{aligned}\sigma_{\sigma_{1,2}}{}^2 = {} & f_0{}^2\sigma_{\nu_0}{}^2 + f_1{}^2\sigma_{\nu_1}{}^2 + f_2{}^2\sigma_{\nu_2}{}^2 + f_3{}^2\sigma_{\nu_3}{}^2 \\ & + 2[f_0 f_1 r_{01}\sigma_{\nu_0}\sigma_{\nu_1} + f_0 f_2 r_{02}\sigma_{\nu_0}\sigma_{\nu_2} + f_0 f_3 r_{03}\sigma_{\nu_0}\sigma_{\nu_3} \\ & + f_1 f_2 r_{12}\sigma_{\nu_1}\sigma_{\nu_2} + f_1 f_3 r_{13}\sigma_{\nu_1}\sigma_{\nu_3} + f_2 f_3 r_{23}\sigma_{\nu_2}\sigma_{\nu_3}]\end{aligned} \tag{15}$$

The quantities σ_{ν_i} and r_{ij} required for calculation of the errors ac-

* In this discussion, expected error and probable error, which have exact definitions, are used interchangeably Since they are constant (arbitrary) fractions of the standard deviation, the latter is the quantity actually calculated.

[12] **Pearson, K.** Biometrika **2**: 273. 1902.

cording to this equation are found by the same process. Thus,

$$\nu_k = \Sigma x_i^k y_i$$
$$d\nu_k = \Sigma x_i^k dy_i$$
$$\sigma_{\nu_k}^2 = \Sigma x_i^{2k} \sigma_{y_i}^2 + \Sigma x_i^k x_j^k r_{ij} \sigma_{y_i} \sigma_{y_j}$$

and

$$r_{kl} \sigma_{\nu_k} \sigma_{\nu_l} = \Sigma x_i^{k+l} \sigma_{y_i}^2 + \Sigma x_i^k x_j^l r_{ij} \sigma_{y_i} \sigma_{y_j}. \tag{16}$$

Now, since the error in y is independent of y, if we consider a deviation δy_i in the ordinate y_i, it is equally likely that the deviation in another ordinate y_j, δy_j, will be positive or negative. Therefore, the correlation coefficient, defined as

$$r_{ij} = \frac{1}{n} \frac{\Sigma \delta y_i \delta y_j}{\sigma_{y_i} \sigma_{y_j}}$$

will be zero, and we have for the final expressions

$$\sigma_{\nu_k}^2 = \Sigma x_i^{2k} \sigma_{y_i}^2 = a^2 \Sigma x^{2k}$$

and

$$r_{kl} \sigma_{\nu_k} \sigma_{\nu_l} = a^2 \Sigma x^{k+l} \tag{17}$$

where a is the (constant) standard deviation in y_i.

Assuming arbitrarily that $a = 1$, corresponding to a one per cent error in the maximum ordinate of the synthetic curve, and using Equations (12), (16), and (17), we obtain

$$\sigma_1 = 8.79 \pm 1.27$$
$$\sigma_2 = 4.33 \pm 4.30$$

The correct values for σ_1 and σ_2 are 9.808 and 5.371, respectively. The deviation of the area of the traced curve from the correct area was 0.75%, and those of the first and second moments correspondingly small, indicating that the value of a assumed is roughly correct.

It appears, then, that great caution must be exercised in applying this suggested method of analysis to the results of ordinary experiments. The synthetic curve used here represents the most favorable choice for a system having the assumed values of σ_1 and σ_2 (equal concentrations of the two solutes) and, obviously, has a considerably lower experimental error than is met with in practice. We have not calculated the error involved in the analysis according to Pearson's procedure, but it will be considerably greater than that discussed above, since it involves moments up to the sixth. Gralén's characteristic constant, on the other hand, will have a relatively smaller expected error. However, some of the experimental results referred to in Part II indicate that the funda-

mental assumption that each species in the mixture diffuses independently is undoubtedly false for some systems.

The above results give point to the observation that the establishment of refractometric methods in routine operations for the characterization of proteins has led, on occasion, to a tendency to ascribe too great accuracy to the results of a single experiment. Estimates of the precision of published results have sometimes been made on the basis of the accuracy with which a comparator vernier can be read, rather than from the agreement between several results for the same system. The need for this second type of comparison may be illustrated by the careful work of L. S. Moody and L. E. Moody, some of whose results are given in TABLE 1 (Part II), and that of N. V. Hakala, from whose dissertation TABLE 7 is taken. In a recent paper by Lauffer,[13] the results of one experiment on tobacco mosaic virus protein are given. Although the maximum variation is fairly large, the use of eight curves representing eight observations in this one experiment gives mean values whose precision is satisfactory:

$$D \times 10^7 \ (0.2^\circ \text{ C}) = 0.262 \pm 0.014 \text{ (height and area)}$$
$$0.241 \pm 0.026 \text{ (second moment)}$$

It is evident that, in order to obtain a precise value for the diffusion constant of even a homogeneous system of spherical molecules, it is necessary to obtain several curves for an experiment, and it is desirable to take the mean of the results of several experiments.

It may be noted that Equations (17) indicate that a greater precision should be obtained when Equation (4a) is used to calculate D, than when Equation (4b) is used, so that the present trend toward use of the height and area method, rather than the customary second moment method, should give results which are in better agreement.

In view of the discussion above, it appears that, in most work, as good an estimate of the diffusion constant of a material known to contain some impurity will be obtained, by using the ordinary procedures, as by applying Equation (12), when diffusion experiments have given the only available data. There are undoubtedly situations where the analysis will be of value when independent evidence is available. If velocity ultracentrifuge or electrophoresis experiments have given a good estimate of the relative concentrations, and if the diffusion constant of one component is known, it should be possible to calculate that of the other. Such calculations would be particularly useful for rapid preliminary characterization of substances obtained in fractionation of

---

[13] Lauffer, M. A. J. Am. Chem. Soc. 66: 1188. 1944.

natural materials, or in the degradation of a known material, when some of the components can be readily obtained in pure form.

It has been frequently observed that the deviation of the experimental curve from the normal, for the diffusion of a mixture, is greatest at the maximum ordinate, and it has been suggested that this deviation could be used to estimate the presence of impurities in a solute. As Neurath[11] has pointed out, the practical application of this depends on the choice of the normal curve used for comparison. The usual one (Equation (6)) has the same area and standard deviation as the experimental curve. It is easily shown that, for this, the deviation at the maximum ordinate is insensitive to the relative concentrations over a wide range. Other possible choices require some knowledge of the properties of one of the components, and, if this is available, quantitative calculations can be made.

The only example of a diffusion curve which shows a negative deviation from the normal at the maximum ordinate is given in Gralén's dissertation.[9b] The curve illustrated is for the diffusion of a probably degraded cellulose. There seems to be no adequate explanation for this observation.

(B) Asymmetrical Deviations from the Normal Curve

Most of the deviations from normality observed in the diffusion of proteins are of the type discussed under Section (A). For some proteins, which are known to have extremely elongated molecules in solution, it is found that, even at the lowest concentrations now experimentally possible, the curves are skewed.[11] Similar results have been obtained in experiments with a few synthetic high polymers.[14] This effect is particularly pronounced in the diffusion of cellulose and its derivatives.[9b] It is obvious, in such cases, that Fick's law no longer applies and other means of interpreting the experimental results must be used.

The lack of published data on the appearance of skewed curves is undoubtedly due to the fact that little use of the velocity ultracentrifuge has been made in the study of high polymers, and even in such papers as those of Signer and Gross[15] on the sedimentation behavior of some polystyrenes, there are given few data from diffusion experiments. Nevertheless, the possibility of making size distribution analyses from sedimentation studies is indicated by these results, and also by those

---

[14] Spaulding, [illegible] Unpublished; Bevilacqua, E. M. Unpublished.
[15] Signer, R., [illegible] Gross. Helv. Chim. Acta 17: 726. 1934. Signer, R. Trans. Faraday Soc. [illegible] Symposium on Polymerization. 1936.

of Bridgman.[16] Thus, the results of the necessary diffusion measurements should be subject to unequivocal interpretation.

An obvious attack on this problem is the development of methods for measuring diffusion constants under conditions where Fick's law does apply. Although there does not appear to be any immediate prospect of an experimental technique which will reduce the minimum concentration difference required by present apparatus, it may be possible to approach ideal diffusion behavior by the proper choice of solvent for some polymers. Theoretical considerations and experimental evidence[17] show that, in solvents in which the polymer is just soluble, the osmotic behavior is much more nearly ideal than in good solvents. Diffusion experiments in which such solvents are used should also show more nearly ideal behavior.

If the diffusion constant is a function of concentration, Fick's law may be modified

$$\frac{\partial c}{\partial t} = \frac{\partial}{\partial x}\left(D\frac{\partial c}{\partial x}\right) \qquad (18)$$

No closed solution of this equation is available. Boltzmann,[18] however, showed that it may be transformed into an equation in a single variable λ, by setting

$$\lambda = \frac{x}{t^{1/2}}$$

from which

$$-\frac{\lambda}{2}\frac{dc}{d\lambda} = \frac{d}{d\lambda}\left(D\frac{dc}{d\lambda}\right) \qquad (19)$$

It has been experimentally found that the concentration is a function of $x/t^{1/2}$ for many systems,[19, 20, 21] and, with this condition fulfilled, values of D may be calculated by graphical or numerical integration. The results of Gralén[9] and of Beckmann and Rosenberg[21] show that, for several polymers, D is a linear function of concentration. These are the first calculations of this kind from the results of experiments using refractometric methods, and it is evident that the method of calculation developed in these papers gives the correct value of D (D_0) to use in calculating molecular weights.

16 **Bridgman, W. B.** J. Am. Chem. Soc. **64**: 2349. 1942.
17 **Flory, P. J.**, J. Am. Chem. Soc. **65**: 372. 1943.
18 **Boltzmann, L.**, Wied. Ann. **53**: 959. 1894.
19 **Barrer, R.**, Diffusion in and through Solids. The MacMillan Company. New York. 1941.
20 **Eversole, W. G., J. D. Peterson, & H. M. Kinsvater.** J. Phys. Chem **45**: 1398. 1941; see also references [6, 7].
21 **Beckman, C. O., & J. L. Rosenberg.** Ann. N. Y. Acad. Sci. **46** (5): 329. 1945.

(C) Other Abnormalities

The discussion of Section (A) rests entirely on the assumption that the various molecular species present in the diffusing material act independently. Bridgman[15] has found that several glycogen preparations give diffusion curves which coincide with the normal curve, although they are known to be heterogeneous with respect to molecular size. Two explanations of this, other than the conclusion that the above assumption is not valid, have been offered. It may be that the range of sizes present is so narrow that the deviations are within experimental error, or that all molecular sizes have the same diffusion constant. The latter explanation does not seem probable, since the diffusion constant should change at a rate inversely proportional to some power of the molecular weight ($M^{-1/3}$ for spheres, faster for asymmetrical molecules[22]). This, together with the range of molecular sizes in the samples studied by Bridgman, makes the first explanation also seem unlikely. A final decision must await an actual fractionation of the material.

A more striking example of the failure of the fundamental assumption is offered by the experiment of Gralén,[9] referred to earlier, since it is impossible for negative deviations from the normal curve at the maximum ordinate to occur if each species is kinetically independent.

The final abnormality discussed here is not of the types with which we are primarily concerned, but it is apparently widely observed Neurath,[11] for example, reports that approximately one-half the experiments carried out at the Duke University School of Medicine give values of D which drift downward with time. Similar effects have been reported elsewhere and observed in this Laboratory. We have found that, in many cases, this is simply due to an unavoidable error introduced by the experimental technique. The boundary formed at the beginning of the experiment is not infinitely sharp, as is required for the application of Equation (2). It is readily seen that, if we assume the time of formation of the boundary is a short time after zero time, plotting the calculated values of D against the reciprocal of the time from formation of the boundary should give a linear relation. This is in agreement with the usually observed variation. The effect is most pronounced in diffusion cells of the Svedberg type and least pronounced in cells using a sheared boundary. Some confirmation for this explanation is given by a series of experiments in a Svedberg type cell in which a compensator driven by a constant hydrostatic head was used to shift the boundary into the optical path of the camera, in place of the pre-

[22] Lundgren, H. P., & J. W. Williams. J. Phys. Chem. 43: 988. 1939.

viously used hand-operated compensator. Values of D for the same solutions showed a smaller change with time although the extrapolated values coincided.[14]

In summary, we may say that there appears to be little justification for modification of the present methods of calculation of diffusion constants in order to account for symmetrical deviations from the normal curve For asymmetrical deviations, the type of calculation developed by Beckmann and Rosenberg may be used if a solvent cannot be found in which the solute shows ideal behavior.

PART II. PRACTICE

Where molecular weight data are available, values of the diffusion constant D may be used to estimate the frictional ratio, and so molecular dimensions. The frictional ratio f/f_0 is obtained from the relations

$$f_0 = 6\pi\eta N \left(\frac{3Mv}{4\pi N}\right)^{1/3} \tag{20}$$

and

$$f = \frac{RT}{D} \quad \text{or} \quad \frac{M(1 - v\rho)}{s} \tag{21}$$

If the deviation of f/f_0 from one is ascribed to shape, the formula of Perrin[23] or of Herzog, Illig and Kudar[24] may be used to calculate the ratio of length to width of the molecule. Thus

$$\frac{f}{f_0} = \frac{[1 - (a/b)^2]^{1/2}}{(a/b)^{2/3} \ln \dfrac{1 + (1 - (a/b)^2)^{1/2}}{a/b}} \tag{22}$$

where a/b = ratio of minor to major axis.

Succeeding paragraphs review selected diffusion constant data illustrating some of the points made in preceding sections.

Proteins and Derived Substances

Amandin and edestin, two seed globulins, are proteins whose molecular kinetic constants have been reported in the literature. A repetition of the work has revealed that a correction of these values is necessary. The diffusion constant for edestin was reported to be 3.93×10^{-7} cm^2/sec. on the basis of one experiment.[25] A series of experiments were made[26] on a preparation of this protein which sedimentation velocity

[23] Perrin, F. J. phys. radium (7) 7: 1. 1936.
[24] Herzog, R. O., R. Illig, & H. Kudar. Z. physik. Chem. A167: 329. 1934.
[25] Polson, A., Kolloid Z. 87: 149. 1939.
[26] Moody, L. S., Dissertation. University of Wisconsin. 1944.

experiments had shown to be essentially monodisperse. The results of ten experiments, calculated by the methods of height and area, D_A, and of moments, D_m, are given in TABLE 1. Each tabulated value is the average of at least four values for one experiment.

TABLE 1
DIFFUSION CONSTANT DATA FOR EDESTIN AT 20°

Protein Conc.	$D_m \times 10^7$ cm²/sec.	$D_A \times 10^7$ cm²/sec.
1.63	3.27	3.30
.95	3.06	3.11
.76	3.25	3.16
.92	3.18	3.19
1.05	3.23	3.22
.62	3.17	3.05
1.63	3.14	3.17
.76	3.04	3.19
——	3.24	3.21
1.05	3.30	3.18
Average.	3.19	3.17
Average deviation.	0.07	0.05
Standard deviation.	0.08	0.06

It is interesting to note that the probable error is smaller when the height and area method is used.

The diffusion constant for amandin was also redetermined. On the basis of fifteen diffusion experiments, the value $D_{20} = 3.45 \times 10^{-7}$ cm²/sec. was adopted, as compared to the value 3.62×10^{-7} cm²/sec. given by Svedberg and Pedersen.[27]

The study of the diffusion behavior of the seed globulins presented no unusual experimental difficulties. The diffusion curves all were normal. They form an excellent illustration of the precision which may be expected in the characterization of pure materials of relatively simple molecular kinetic behavior.

Asymmetrical proteins have given more trouble. For example, Carter[28] has studied the properties of calf thymus nucleoprotein. The nucleoprotein proved to be essentially monodisperse in a series of sedimentation velocity experiments. The diffusion curves were skewed with the result that calculation by the usual methods showed a difference between D_A and D_m, in spite of the fact that the protein was homogeneous. The results of some experiments are given in TABLE 2.

[27] Svedberg, T., & K. O. Pedersen. The Ultracentrifuge. The Clarendon Press Oxford. 1940.
[28] Carter, R. O. J. Am. Chem. Soc. 63: 1960. 1941.

TABLE 2
DIFFUSION OF CALF THYMUS NUCLEOHISTONE AT 25°

	Conc. (g/100 cc)	Solvent	$D_A \times 10^7$ cm²/sec.	$D_m \times 10^7$ cm²/sec.	D_m/D_A	f/f_0	M
(1)	0.51	Phosphate buffer pH = 6.4 0.855 M NaCl	0.83	1.39	1.67		
(2)	0.20	Same as above	0.64	1.12	1.75		
(3)	0.46	Phosphate buffer pH = 7.0	1.03	1.07	1.04	2.5	2,300,000

Since the molecular weight by sedimentation equilibrium measurement is 2,000,000, the true value of the diffusion constant cannot be greatly different from 1×10^{-7} cm²/sec.

Several nucleic acid preparations, one derived from the preceding nucleoprotein, have been investigated by Tennent and Vilbrandt[29] with results shown in TABLE 3. These samples also appeared to be homogeneous in sedimentation experiments, but in no case did the diffusion curve coincide with the normal curve. The diffusion constants reported were calculated by the method of moments.

TABLE 3
PHYSICAL PROPERTIES OF NUCLEIC ACIDS AT 25°

Sample	$D_m \times 10^7$ cm²/sec.	f/f_0	b/a	s (in svedbergs)	M
Sodium thymonucleate-1 (STN-1)	0.61	8.0	400	6.4	580,000
STN-2	1.0	5.3	170	8.0	430,000
STN-3	0.95	5.6	200	7.8	450,000
Thymonucleic acid-1	21.5	1.1	3	1.8	4,800
Pancreas polynucleotide	18.4	1.2	4	2.4	6,700

Another example of the effect of the shape of a molecule on its diffusion behavior is shown by the virus proteins. The tomato bushy stunt virus, the rabbit papilloma virus, and the tobacco mosaic virus proteins all have very large molecular weights, but vary in shape from a spherical molecule to a very elongated one. Spherical bushy stunt virus showed normal diffusion at concentrations below 1.2 per cent,[30] while the asymmetric tobacco mosaic virus showed anomalous diffusion be-

[29] Tennent, H. G., & C. F. Vilbrandt. J. Am. Chem. Soc. 65: 424. 1943.
[30] Neurath, H., & G. R. Cooper. J. Biol. Chem. 135: 455. 1940.

havior at concentrations of 0.3 per cent.[31] The intermediate rabbit papilloma virus showed ideal diffusion behavior below 0.3 per cent.[32]

TABLE 4

DIFFUSION CONSTANT DATA FOR THE VIRUS PROTEINS AT 20°

Protein	Conc.	$D \times 10^7$ cm²/sec.	f/f_0	M(s and D)
Tomato bushy stunt virus	0.2 0.4	1.2	1.27	10,600,000
Rabbit papilloma virus	0.2	0.59	1.49	47,000,000
Tobacco mosaic virus	0.2 0.2	0.3 (1938) 0.53 (1944)	2.5	60,000,000 31,600,000

One must conclude from this table that the diffusion constant for native tobacco mosaic virus protein is still not fully established. The 1944 value is taken from the recent article by Lauffer.[13]

Polysaccharides

Although work on polysaccharides has been rather limited, a number of these compounds have been studied in this Laboratory.

The polysaccharides from filtrates of human tubercle bacillus have been studied by Seibert, Pedersen, and Tiselius,[33] Tennent and Watson,[34] and Bevilacqua[35] (TABLE 5). In all the experiments reported diffusion was normal. The human tubercle bacillus polysaccharide prepared by Bevilacqua[35] is considerably different from the other samples, and is apparently the least degraded. The difference between preparations 1 and 2 is not considered significant.

Tennent and Watson[34] also studied polysaccharides from bovine tubercle bacilli, but these were too hetergeneous to give significant diffusion constants.

The characteristics of glycogen included in TABLE 5 were obtained by Bridgman,[15] whose results were mentioned in Part I.

The pectins investigated by Säverborn[36] are of interest because of their unusually high frictional ratios, although the molecular weights are not extremely high. The diffusion experiments showed curves which were extremely asymmetrical, even at concentrations as low as 0.16%,

31 Neurath, H., & A. M. Saum. J. Biol. Chem. 126: 435. 1938

32 Neurath, H., G. R. Cooper, D. G. Sharp, A. R. Taylor, D. Beard, & J. W. Beard. J. Biol. Chem. 140: 293. 1941.

33 Seibert, F. B., K. O. Pedersen, & A. Tiselius. J. Exptl. Med. 68: 413. 1938.

34 Tennent, D. M., & D. W. Watson. Jour. Immun. 45: 179. 1942.

35 Bevilacqua, E. B. Dissertation. University of Wisconsin 1944.

36 Säverborn, S. Kolloid Z. 90: 41. 1940.

so that the variation of diffusion constant with concentration must be quite large.

TABLE 5

DIFFUSION CONSTANT AND OTHER PHYSICAL DATA FOR POLYSACCHARIDES AT 20°

Compound	$D \times 10^7$ cm²/sec.	f/f_0	b/a	s (in svedbergs)	M	Reference
Tubercle bacillus polysaccharide						
Human	11.0	1.5		1.8	9,000	33
Human	12.4	1.40	7.4	1.39	7,200	34
Human (1)	7 0	1.71	13	2 0	23,000	35
(2)	7.6	1.71	13	1.7	18,000	35
Avian	13.6			1.54	7,300	34
Leprosy bacillus	24.9	0.99	1.0	0.97	2,500	34
Glycogen	1.1	1.90	18	65	4,100,000	15
	1.1	1.94		61	3,900,000	
	1.1	1.76		82	5,200,000	
	1.1	1 83		73	4,600,000	
Pectins						
Apple (1)	0.83	8 3		2 8	117,000	36
(2)	1.4	5 1		2.3	99,000	
Currant	2.85	3.7		2.0	33,000	
Citrus (albedo)	0.65	8 0		4.0	271,000	

Cellulose and Cellulose Derivatives

Polson[37] has studied the diffusion of some cellulose derivatives in detail. His samples were fractionated materials, the molecular weights of which had already been determined by Signer and co-workers.

The methyl celluloses were investigated in aqueous sodium chloride solutions in concentrations between 0.5 and 1.0%. Below 0.5%, the experimental accuracy was too small; above 1.0%, the diffusion curves were anomalous. Within these limits, the curves showed fair agreement with the normal. A methyl cellulose of molecular weight 100,000 was also studied, but gave curves which were too skewed to give a significant diffusion constant.

The molecular weights calculated from diffusion and sedimentation agree very well with the results reported by Signer. It is interesting to note that the sedimentation constant is essentially unchanged over this range of molecular weights, and the diffusion constant, which is a

[37] Polson, A. Kolloid Z 83: 172. 1938.

measure of the length of an asymmetrical molecule, accounts for the variation in molecular weight.

TABLE 6
DIFFUSION AND OTHER RELATED CONSTANTS FOR METHYL CELLULOSE AT 20°

Molecular Weight by Signer	f/f_0	b/a	s (in svedbergs)	$D \times 10^7$ cm²/sec	M (s and D)
38,100	4.5	139	0.89	2.47	33,000
24,300	3.77	109	0.79	3.05	22,600
14,100	3.04	77	0.83	4.45	14,200

The recent dissertation by Gralén[9b] contains a very extensive study of the sedimentation and diffusion of cellulose and cellulose derivatives. Native fiber and wood celluloses, cellulose nitrates, sodium cellulose xanthate, and sodium cellulose glycolate were studied.

The experiments with cellulose in cuprammonium showed behavior which was far from ideal. Although it was possible to calculate diffusion constants from the skewed curves according to the procedure based on the Boltzmann equation, the significance of the results is not clear. It was found that diffusion in the cuprammonium-cellulose system was apparently retarded for long periods after formation of the boundary and the induction periods were reproducible. The data given by Gralén were calculated by using the time from the end of the induction period as zero time. It was also found that the ratio of the diffusion constants calculated by the second moment and by height and area methods, D_m/D_A, was frequently less than one, although the materials were all heterogeneous.

The diffusion of the cellulose derivatives gave evidence of polydispersity and of concentration dependence of the diffusion constants, but did not show the anomalies referred to above and in Part I.

Sodium Lauryl Sulfate

The diffusion constant data for purified sodium lauryl sulfate[38] included in TABLE 7 are presented to show an additional example of the necessity to use the results of a number of experiments in order to obtain a mean value of the diffusion constant which is satisfactorily precise.

[38] Hakala, N. V., Dissertation. University of Wisconsin 1943.

TABLE 7
VARIATION OF DIFFUSION CONSTANT OF SODIUM LAURYL SULFATE IN NaCl SOLUTION AT 20°*

Conc. SLS (N)	Conc. NaCl (N)	$D_m \times 10^7$ cm²/sec.	$D_A \times 10^7$ cm²/sec.
0.05	0.20	8.7	8.8
		7.8	8.1
		8.1	8.5
		8.6	8.5
		9.1	9.1
0.04	0.20	8.1	8.9
		9.4	9.3
		8.3	9.0
0.03	0.20	9.2	9.2
		9.2	9.9
		8.6	8.7
0.02	0.20	8.8	8.5
		8.8	8.8
0.01	0.20	9.3	8.6
		9.9	8.3
		9.7	9.8
0.04	0.15	7.8	9.0
		9.0	9.5
0.04	0.10	8.0	9.1
		8.5	8.9
0.04	0.05	9.2	8.9
		8.7	8.7
0.03	0.15	8.7	8.5
		8.9	8.5
0.02	0.10	9 1	8.8
		8.5	8.6
Average		8.77	8.87
Average deviation		0.43	0.33
Standard deviation (σ)**		0.53	0.43
DIFFERENTIAL DIFFUSIONS			
0.04/0.02	0.20	8.3	8.9
		8.9	9.4
0.03/0.01	0.20	8.7	8.5
		8.8	8.4

* Each value given is the average of one experiment.

** $\sigma = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n}}$

The authors wish to acknowledge their gratitude to Drs. N. V. Hakala, L. S. Moody, and L. E. Moody, for permission to use some of their previously unpublished data. In addition, some of the ideas considered in this report have evolved as a result of discussions not only with them, but with other former members of this Laboratory.

The experimental work was carried out with the aid of generous grants from the Wisconsin Alumni Research Foundation.

THE EFFECTS OF CONCENTRATION AND POLYDISPERSITY ON THE DIFFUSION COEFFICIENTS OF HIGH POLYMERS

BY CHARLES O. BECKMANN AND JEROME L. ROSENBERG
Department of Chemistry, Columbia University, New York, N. Y

INTRODUCTION

The differential equation of diffusion,

$$\frac{\partial c}{\partial t} = \frac{\partial}{\partial x}\left(D\frac{\partial c}{\partial x}\right) \qquad (1)$$

in which c is the concentration of the solute, x is the distance along the direction of diffusion, t is the time and D is the diffusion coefficient, cannot be integrated in closed form if D is an arbitrary function of the concentration, c. By taking into account that c is a function of $z = x/\sqrt{t}$ only, the equation may also be written in the form

$$\frac{d}{dz}\left(D\frac{dc}{dz}\right) = -\frac{z}{2}\frac{dc}{dz} \qquad (2)$$

Boltzmann[1] has shown that if c is known experimentally as a function of z, one may obtain from this equation a value of D for every value of c from zero to the initial concentration, c_0, of the solution. On integration of Equation (2), at constant t, one obtains

$$D(c) = -\frac{1}{2t}\frac{dx}{dc}\int_0^c x\,dc = -\frac{1}{2t}\frac{dx}{dc}\int_{-\infty}^{x} x\left(\frac{dc}{dx}\right)dx \qquad (3)$$

Because dc/dx is proportional to the refractive index gradient dn/dx in dilute solutions, all optical methods which measure this gradient as a function of the position, x, in the diffusion cell are particularly suited for the evaluation of the diffusion coefficient $D(c)$ from Equation (3).

Lamm,[2] in his extensive analysis of the diffusion problem, has shown that there are many advantages to be realized by transforming the experimental data obtained at various times of diffusion to a set of normal coordinates which eliminate time, diffusion coefficients, concentration, and geometry of optical system as variables. In such coordinates, the experimental points obtained at different times, t, will all fall on the same curve for a perfect experiment. In addition, therefore, to yielding a convenient method of calculation, one has available a sensitive

[1] Boltzman, L. Ann. d. Physik, 53: 959. 1894.
[2] Lamm, Ole. Nova Acta Reg. Soc. Upsal. IV, 10 (6). 1937.

test for the reliability of data obtained at different times of diffusion.

One proceeds by first evaluating an average diffusion coefficient, $D_{2,0}$, from the second and zero moments of the original experimental curves and the equation

$$\frac{\int_{-\infty}^{+\infty} x^2 \left(\frac{dc}{dx}\right) dx}{\int_{-\infty}^{+\infty} \left(\frac{dc}{dx}\right) dx} = 2D_{2,0}\,t = \sigma^2 \tag{4}$$

Gralén[3] has shown that the $D_{2,0}$ so calculated is a weight average diffusion coefficient for systems that are polydisperse. Now, if one sets

$$X = \frac{x}{2\sigma} \quad \text{and} \quad Y = \frac{5\sigma}{\Phi}\left(\frac{dn}{dx}\right)\alpha \tag{5}$$

where 2 and 5 are arbitrary constants chosen for convenience of plotting, α is a constant of the optical system connecting the measurement on the photographic plate with the refractive index gradient in the cell, Φ is the area of the $\alpha\left(\frac{dn}{dx}\right)$ versus x curve, and σ is defined by Equation (4), one converts Equation (3) into

$$D(c) = -\frac{4D_{2,0}}{Y}\int_{-\infty}^{X} XY\,dX \tag{6}$$

For the case of a mono-disperse solute with a constant diffusion coefficient, the equation for the dn/dx versus x curve in normal coordinates becomes

$$Y^{o} = \frac{5}{\sqrt{2\pi}}\,e^{-2X^2} \tag{7}$$

This will be called the ideal diffusion curve and its Y ordinate will be characterized by a superscript zero.

CONCENTRATION EFFECTS

When the diffusion coefficient is concentration dependent, the experimental curve and the normalized curve will be skew with the maximum no longer coincident with the position of the original boundary between solution and pure solvent (See FIGURE 1). If the diffusion coefficient increases with increasing concentration (as is usually the case), the maximum is displaced to the pure solvent side (negative values of X in the figures) of the boundary. And if the diffusion coefficient decreases with increasing concentration, the maximum is dis-

[3] Gralén, N. Koll. Zeits. 95: 188. 1941

placed to the solution side. Gralén[4] has shown that, if the diffusion coefficient varies linearly with the concentration according to

$$D = D_0(1 + kc) \tag{8}$$

one may calculate the value of k from the ratio of the values of x and dn/dx, at the maximum point of the curve. In the normal coordinate

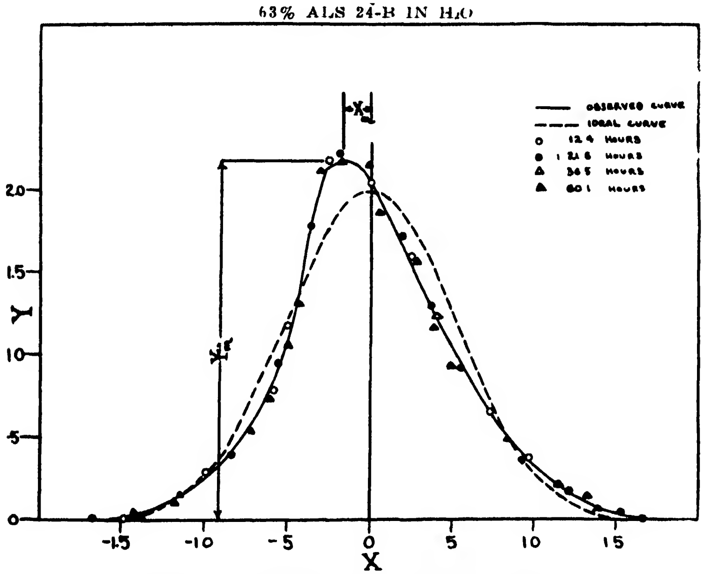


FIGURE 1 Normalized skew curve plotted with ideal curve

system defined by Equation (5), his equation becomes simply

$$k = -\frac{10}{c_0}\frac{D_{2,0}}{\bar{D}_0}\frac{X_{max}}{Y_{max}} \tag{9}$$

and the relation between $D_{2,0}$ and D_0 is given by

$$D_0 = D_{2,0}\left(1 + \frac{5X_{max}}{Y_{max}}\right) \tag{10}$$

In these equations, the subscript "max." refers to the values of the coordinates at the maximum of the curve. This simple method has a limitation, however, in the *a priori* assumption that the concentration dependence of the diffusion coefficient is given by Equation (8). To

[4] Gralén, N. Inaugural Dissertation Upsala 1944

obtain values of $D = f(c)$ that are free of such limitation, one must resort to a numerical or graphical integration of Equation (6).

To simplify the process of graphical integration, it has been found convenient to construct a drawing curve (made of brass) of the ideal diffusion curve, Equation (7), and to use this to draw the ideal curve on the same graph as the normal curve (See FIGURE 1). By measuring off differences between the Y ordinates of the two curves at various values of X, one may calculate $D(c)$ from the equation

$$D(c) = D_{2,0}\frac{Y^o}{Y}\left[1 - \frac{4}{Y^o}\int_{-\infty}^{X} X\Delta Y dX\right] \tag{11}$$

where $\Delta Y = Y - Y^o$. In practice, the integral is, of course, replaced by a sum, the interval of X taken being ordinarily 0.1, and the sum evaluated in the usual way by use of the trapezoidal rule. The result expresses $D(c)$ as a function of X. To convert to a function of concentration, c, one makes use of the relation

$$\frac{c}{c_0} = \frac{2}{5}\int_{-\infty}^{X} Y dX \tag{12}$$

and treats the integral as a sum in the manner described for Equation (11). Or, one may employ the value of ΔY, already recorded, and the equation

$$\frac{c}{c_0} = \frac{1}{2}\left[1 + \frac{2}{\sqrt{\pi}}\int_{0}^{X\sqrt{2}} e^{-x^2}dx + \frac{4}{5}\int_{-\infty}^{X} \Delta Y dX\right] \tag{13}$$

which is easily obtained by setting $Y = Y^o + \Delta Y$ and substituting Equation (7) for Y^o. The value of the probability integral must be added for positive values of X and subtracted for negative values At the original boundary, $X = 0$, c is approximately equal to $c_0/2$, for, in all cases studied by us, the value of the last integral in Equation (13) is very small at this point.

The results are most readily expressed by plotting $D/D_{2,0}$ against $2c/c_0$, and some are shown in FIGURE 2. It will be shown in the next section that the curves with minima in the vicinity of $2c/c_0$ equal to unity belong to systems that are polydisperse and have no concentration dependence. The two samples for which straight lines are drawn show a marked dependence. Neglecting the lower parts of the curves (due to polydispersity), it is obvious that one may represent the dependence on concentration by the constant (k Equation (8)). In TABLE 1, the results of a number of experiments are collected (c is expressed in all following tables and figures as grams of solute per 100 g. solu-

tion). With the exception of the last two samples in the series, it is to be noted that k is not very large. At present, insufficient data are at hand to enable one to draw general conclusions about the variance of k

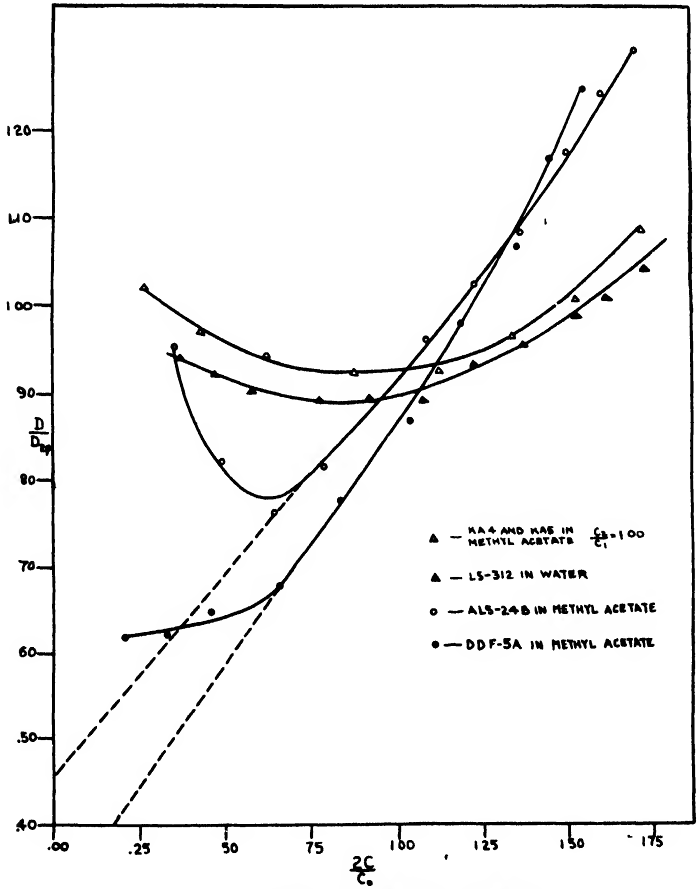


FIGURE 2 Experimental curves based on Equation (11)

with size and shape. The only samples that are at all comparable are the three amylose acetates, KA4, KA3 and KA5. It has been shown that the molecules of these samples are helices of the same diameter but with different lengths In this series, we find the longest molecule

TABLE 1

Sample No.	Material	Solvent	$D_0 \times 10^7$	k	M	Degree of Homogeneity	Probable Shape of Molecules
—	KCl	Water	187	0.0	—	—	—
LS-648 [5]	Acid treated starch	Water	11.5	0.5	5,000	low	?
LS-312 [5]	Acid treated starch	Water	10.0	0.2	10,000	low	?
AS-1R [6]	Ground starch acetate	Methyl Acetate	15.3	0.0	32,000	?	?
KA4 [7]	Potato Amylose Acetate	Methyl Acetate	8.0	0.0	69,000	very high	helical
23B [8]	Cellulose Acetate	Acetone	5.4	0.0	100,000	high	linear
KA3 [7]	Corn Amylose Acetate	Methyl Acetate	6.8	0.0	108,000	very high	helical
KA5 [7]	Tapioca Amylose Acetate	Methyl Acetate	4.2	0.6	151,000	very high	helical
120-2 [9]	Polystyrene	Toluene	3.6	0.1	240,000	high	linear
ALS-24B [5]	Potato Amylose Acetate	Methyl Acetate	2.2	3.4	350,000	low	helical
DDF-5A	Corn Amylopectin Acetate	Methyl Acetate	0.42	3.6	17×10^6	very low	highly branched

with a slight concentration dependence. Gralén,[10] in his extensive study of cellulose and some of its derivatives, finds values of k of higher magnitude. For a group of samples ranging in molecular weight from 37,000 to 400,000, the values of k vary from 0.0 to 2.4; for another group up to 2,500,000, k goes to 7.0; and for one sample of native flax fiber with a molecular weight estimated as 8.2×10^6, k is 22.0.

[5] Moran, F. B. Dissertation. Columbia University. 1945.
[6] Bryce, H. G. Dissertation. Columbia University. 1943.
[7] Dembrow, D. A. Dissertation. Columbia University. 1944.
[8] Badgeley, W. J., & H. Mark. Unpublished data.
[9] Alfrey, T., A. Bartovics, & H. Mark. J. Amer. Chem. Soc. 65: 2319. 1943.
[10] Gralén, N. Inaugural Dissertation. Upsala. 1944.

In general, there is an increase of k with molecular weight, but a quantitative relationship cannot be deduced at the present time.

A complete theory of the diffusion of high polymer molecules in solution would, of course, include an explanation of the concentration effect. In the absence of such a theory, one must limit oneself to an examination of the individual quantities appearing in the equation of Onsager and Fuoss[11]

$$D = \frac{RT}{f}\left(1 + \frac{d \ln \gamma}{d \ln c}\right) \tag{14}$$

which was derived from a general treatment of the diffusion problem. In this equation, f is the frictional coefficient and, γ, the activity coefficient of the solute. Since both f and γ may be concentration dependent, the problem resolves itself into two, the hydrodynamic problem (to determine $f = f(c)$) and the thermodynamic problem (to determine $\gamma = \gamma(c)$).

The latter is more easily analysed since the activity coefficient may be calculated from osmotic pressure data. The osmotic pressure, π, of dilute high polymer solutions may be expressed by

$$\frac{\pi}{c} = \frac{RT}{M} + bc \tag{15}$$

where M is the molecular weight of the solute and b is a positive constant independent of M in a polymer homologous series. From this equation, one can show that

$$\frac{d \ln \gamma}{d \ln c} = \frac{2bM}{RT} c \tag{16}$$

By substituting this value in the Equation (14) of Onsager and Fuoss, one obtains a result that indicates that the diffusion coefficient increases with concentration and that the rate of increase is proportional to the molecular weight. Both conclusions are also deducible from experiment, but the lack of quantitative agreement between calculation and experiment shows that the variation of the frictional coefficient, f, with concentration, is as important a factor as the variation of the activity coefficient, γ.

The great difficulty of the general problem can be localized in the hydrodynamic problem. The problem of the motion of two spheres in a viscous medium has been treated by many workers and the general conclusion drawn is that hydrodynamic interaction always de-

[11] Onsager, L., & R. M. Fuoss. J. Phys. Chem. 81: 158. 1938.

creases the resistance to flow.[12] Extension of the mathematical methods to a many-body problem has not been carried out and is besct with many difficulties. Powell and Eyring[13] have proposed the idea that the frictional coefficient is proportional to the viscosity of the solution. The fact that the diffusion coefficient always varies less with concentration than would be indicated by the thermodynamic factor of Equation (14) alone gives qualitative support to this theory. Onsager[14] in discussing the problem has set

$$f = a\eta \tag{17}$$

where η is the viscosity of the solution and "a" is a proportionality factor which may also be concentration dependent. The variation of "a" with concentration may serve, therefore, as a measure of deviation from the Powell and Eyring theory.

The evaluation of the many effects is possible for two samples for which the necessary data are at hand. In both cases, the diffusion coefficients are not strongly dependent on concentration while the activity coefficients and viscosity are strongly dependent on concentration. From the osmotic pressure data of a well-fractionated sample of cellulose acetate[8] (23B, TABLE 1), the activity coefficient was calculated as a function of concentration. The individual effects may be compared by plotting the functions η/η_0 and $(1 + d \ln \gamma/d \ln c)$ against c (FIGURE 3). On the same graph, are given the ratio of these functions, namely, $(1 + d \ln \gamma/d \ln c)\eta_0/\eta$ (which is also a^D/a_0D_0, if the subscript zero refers to zero concentration) and the measured D/D_0. The lack of coincidence of the two latter curves shows that, in this case, the Powell and Eyring theory is incomplete (i.e. "a" is a function of concentration; in this case, decreasing with increasing concentration).

The osmotic and viscometric data on the second sample, polystyrene (120-2) in toluene, were reported by Alfrey, Bartovics and Mark.[9] These data were combined with our diffusion data on the same sample and plotted in the manner described above (FIGURE 4). In this case, the function, $(1 + d \ln \gamma/d \ln c)\eta_0/\eta$, increases with concentration, even more steeply than does the diffusion coefficient. Thus, the factor "a" increases with concentration, in this case.

It is thus seen that the hydrodynamic factor and the thermodynamic factor approximately cancel one another and qualitatively account for

[12] See, for example, **Oseen, C. W.** Hydrodynamik: 203–208. Akademische Verlagsgesellschaft M. B. H. Leipzig. 1927.
[13] **Powell, R. E., & H. Eyring.** Advances in Colloid Science I: 183. 1942. Interscience Publishers. New York, N. Y
[14] **Onsager, L.** Ann. N. Y. Acad. Sci. 46 (5): 249 (Cp. Onsager's Equation (29)). 1945.

the fact that diffusion coefficients are not greatly dependent on concentration.

This conclusion has a number of important corollary conclusions with regard to the problem of the variation of the ultracentrifugal sedi-

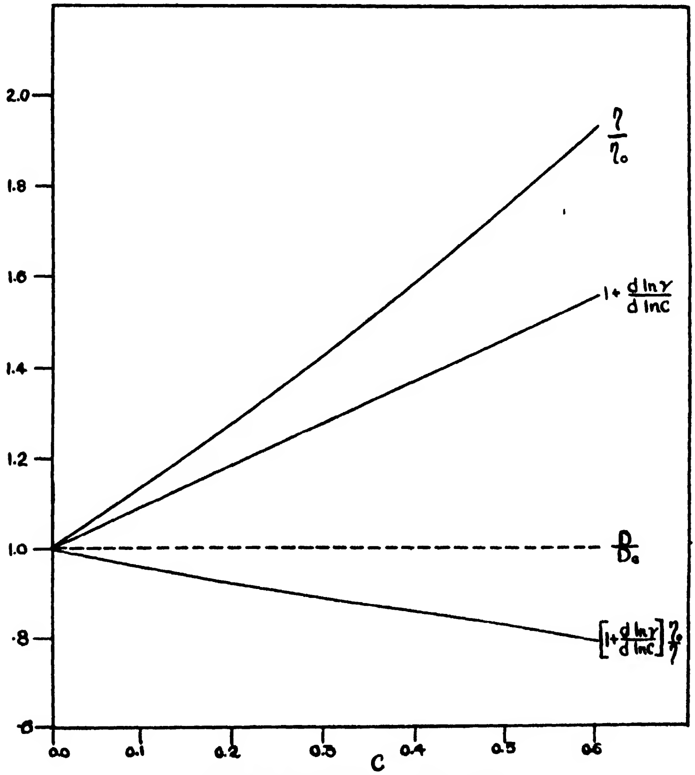


FIGURE 3. 23B Cellulose acetate in acetone.

mentation constant with concentration. Bryce and Beckmann[15] have shown that the sedimentation constant is dependent on concentration according to the equation,

$$s = \frac{M(1 - V\rho)D}{RT\left(1 + \dfrac{d \ln \gamma}{d \ln c}\right)} \tag{18}$$

[15] Bryce, W. G., & C. O. Beckmann. Paper read at the Pittsburgh Meeting of the American Chemical Society. September, 1943. Bryce, W. G. Dissertation. Columbia University. 1943.

where, in general, both D and γ are functions of concentration. From the cases studied by them, they concluded that D is essentially constant and that the greater part of the variation of the sedimentation constant was due to the thermodynamic factor. Lauffer,[16] in a recent

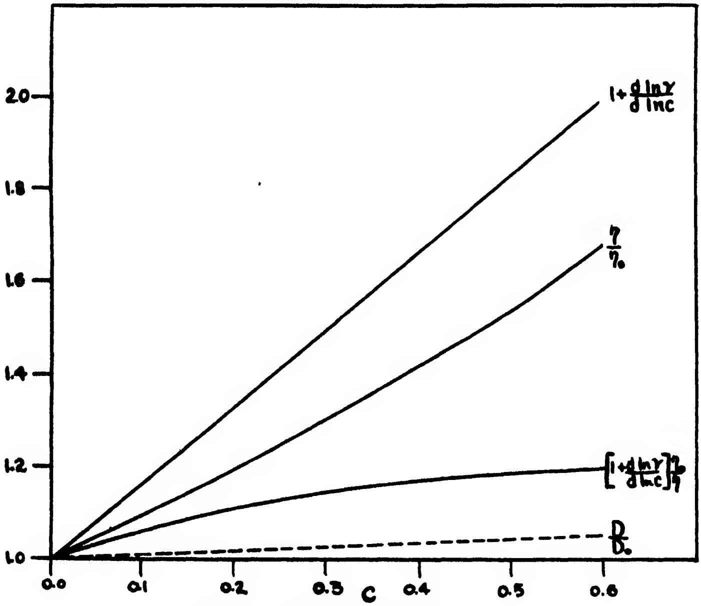


FIGURE 4. 120-2 Polystyrene in toluene.

article, shows that one can account for the effect by correcting for the increase in viscosity of the solution with increasing concentration, i. e.

$$s = \frac{s'\eta_0}{\eta}$$

where s', is the corrected sedimentation constant. It is now obvious that both conclusions are equivalent, provided the Onsager factor "a" is independent of concentration.

Many workers have made the observation that the frictional coefficient, f_D, for the diffusion process and the frictional coefficient, f_s, for the sedimentation process, are not identical. In all cases known to us, the effect of the variation of activity coefficient with concentration

[16] Lauffer, M. A. J. Amer. Chem. Soc. 66: 1195. 1944.

was not considered. Such was the treatment of Beckmann and Landis,[17] which expressed the frictional coefficients explicitly:

$$M = \frac{RTs}{(1 - V\rho)D} \cdot \frac{f_s}{f_D}$$

where f_D was defined by

$$f_D = \frac{RT}{D}$$

By recasting this definition in terms of Equation (14), one is led to

$$M = \frac{RTs\left(1 + \dfrac{d \ln \gamma}{d \ln c}\right)}{(1 - V\rho)D} \cdot \frac{f_s}{f}$$

where f is the frictional coefficient from the Onsager-Fuoss treatment of diffusion.[11] Equality of f_s and f would identify this equation with Equation (18) and would lead to the obvious result

$$\frac{f_s}{f_D} = 1 + \frac{d \ln \gamma}{d \ln c}$$

Neglect of the thermodynamic term would result, in the case of high polymers, in an apparent molecular weight which decreases with increasing concentration. The same error arises in sedimentation equilibrium measurements, if concentration is allowed to replace activity in the equation, as shown by Bryce and Beckmann.[15] The results of Signer and Gross[18] on polystyrene substantiate these conclusions.

THE EFFECTS OF POLYDISPERSITY

The behavior of a polydisperse system with no concentration dependence of diffusion coefficient has been discussed in detail by Lamm. He proved the general proposition that, in such cases, the X coordinate of the maximum of the X, Y curve is not displaced from the position of the original boundary, but the Y coordinate, at this position, is always higher than the Y^0 of the ideal curve. This is illustrated in FIGURE 5, for the hypothetical case of eight components having values of σ ranging from 0.2 to 0.55. A value of Y_{max} greater than Y^0_{max} can always be taken as evidence of polydispersity.

If one naively treats a curve of a polydisperse system as one in which there is only a concentration effect, one arrives at conflicting results when both Equations (9) and (11) are used to evaluate the magnitude of the effect. Because the position of X_{max} is unchanged, Equa-

[17] **Beckmann, C. O., & Q. Landis.** J. Amer. Chem. Soc. 61: 1495. 1939.
[18] **Signer, R., & H. Gross.** Helv. Chim. Acta 17: 59, 335. 1934.

tion (9) leads to $k = 0$, *i.e.*, no concentration dependence. On the other hand, because the X, Y curve of the polydisperse system does not coincide with the ideal curve at all points, one would calculate a concentration dependence from Equation (11). The result, however, is different from that of a monodisperse system (See FIGURE 6). One finds that a plot of $D/D_{s,0}$ versus $2c/c_0$ goes through a minimum at unity on the abscissa scale and is symmetrical about this point. A spurious concentration dependence of this type is easily detectable

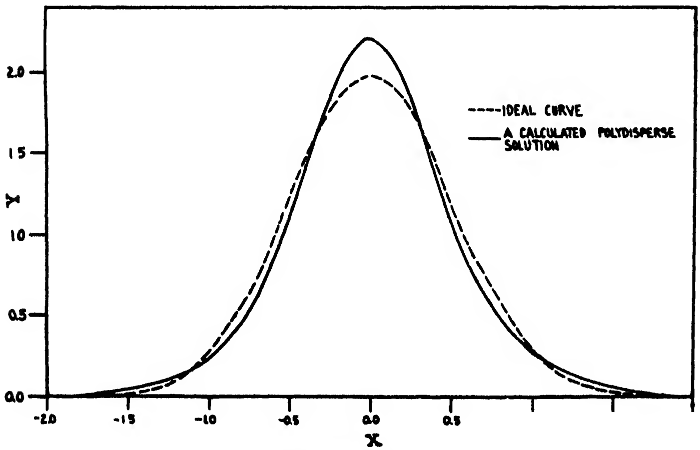


FIGURE 5. Normalized diffusion curves.

The general shape of the curve and the position of the minimum is independent of the original concentration and the spurious character of the curve could be detected by the comparison of two experimental data at two different original concentrations.

When both effects, i.e., concentration and polydispersity, are present together, the resulting curve has a minimum which is displaced to a value of $2c/c_0$ less than one, if the concentration dependence is of the usual kind, increasing D with increasing c. Such cases have been found and are illustrated in FIGURE 2. The polydispersity of these systems has been verified by fractionation and analyses of the fractions. If the two effects are additive, a reasonable assumption, the slope of the line in the vicinity of $c = c_0/2$ gives the value of k characteristic to the system. Values of k obtained this way have been found to be larger than the values of k obtained by Gralén's method. This

is probably due to an increase of the value of Y_{max}, due to polydispersity, although this point has not been proven for systems of this kind.

The relation of the value of the k of a polydisperse system to the values of k of the component fractions presents an interesting problem. An attempt to discover the relationship was made by comparing the

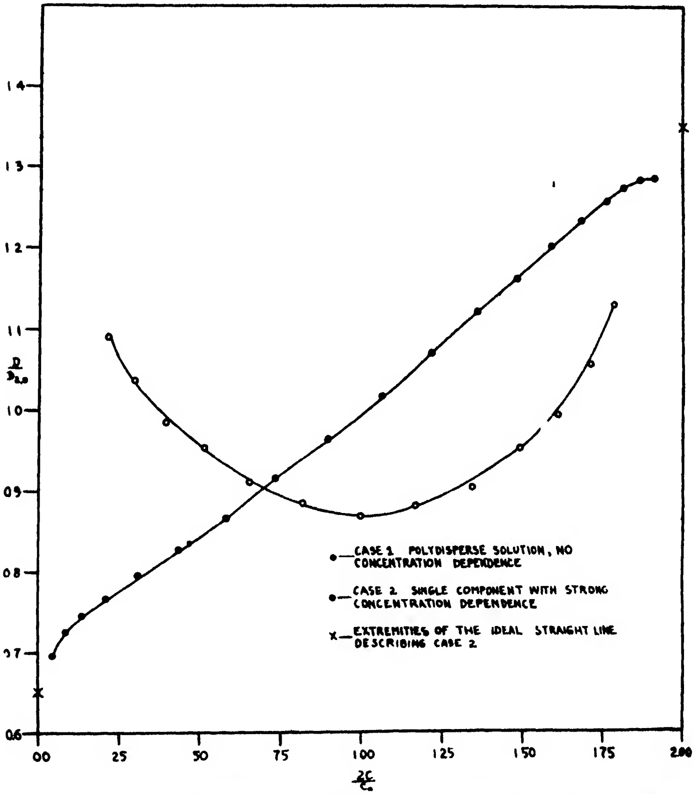


FIGURE 6 Apparent concentration dependence from two calculated diffusion curves

diffusion coefficients of two highly fractionated amylose acetates (KA4 and KA5) with the weight average diffusion constant (given by Equation (4)) of a 50:50 mixture of the two, at the same total concentration (0.44%). One of the amylose acetates showed a concentration effect (KA5, k = 0.6), while the other did not. If, in a mixture of the two, each component exhibited a concentration dependence that was inde-

pendent of the presence of the other, one would obtain the curved dotted line of FIGURE 7. On the other hand, if one assumed that the diffusion coefficient of the one component, which showed a concentration dependence, was dependent on the total concentration of material present, the average diffusion coefficient of mixtures would be given by the straight line. The measured diffusion coefficient falls above both

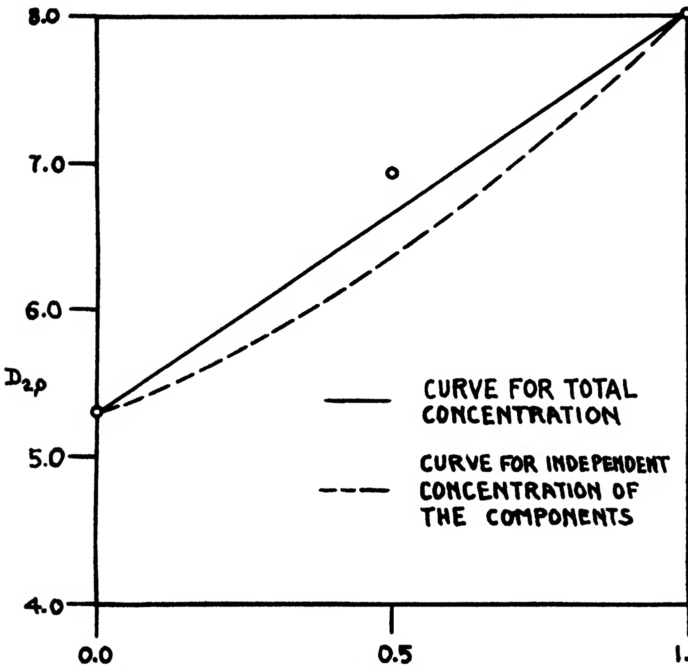


FIGURE 7 KA4 and KA5 in methyl acetate Weight fraction KA4 in solute

these lines, which indicates a third possible situation, namely, that the second component, which, by itself, shows no concentration dependence, acquires one because of interaction with the first component Actually, the experimental error is such that one cannot decide between the second and third possibilities. More conclusive experiments should be performed to determine the nature of the additivity of the individual k-values.

Although curves, such as those shown in FIGURE 2, characterize polydispersity, they cannot be used to measure it. Bevilacqua, Bevilacqua,

Bender & Williams[19] have reviewed the various attempts to express homogeneity in a quantitative fashion based on diffusion alone; but in all cases considered, the curves were completely symmetrical, that is, there was no concentration dependence. Perhaps the simplest criterion of heterogeneity is Y_{max} on the normalized diffusion curve, suggested by Lamm.[2] The greater the value of Y_{max}, the more heterogeneous is the material. The application of this criterion, even to skew curves, is justified because concentration dependence itself does not result in a value of Y_{max} much greater than Y^0_{max}, the value for the ideal curve. Numerical integration of the diffusion equation was carried out for several cases of linear dependence of D on concentration for a monodisperse substance. For values of the ratio, D at c_0, initial solution concentration, to D at infinite dilution, equal to 1.00 (no concentration dependence), 1.50, 2.08, 3.00. The corresponding values of Y_{max} are 2.00, 2.00, 2.03, 2.07.

Another test of polydispersity is possible by comparing the results of a free diffusion experiment with those of an ultracentrifugal experiment. During the process of free diffusion, the original boundary remains fixed during the experiment. In the ultracentrifuge, on the other hand, the diffusion boundary changes with time and is resolved into a series of boundaries, one for each species. The resultant dn/dx versus x curve is, therefore, abnormally wide and low. By the proper adjustment of constants and superposition, the free diffusion curve may be compared with the ultracentrifuge curve. The constants of the two curves may be made in many ways, e.g., one may make the maximum heights equal, or one may make the areas equal. By using the latter method, the curves of FIGURES 8 and 9 were obtained. In FIGURE 8, the two curves are coincident within experimental error. One may conclude that there was no resolution of boundaries in the ultracentrifuge and that the sample was, therefore, monodisperse. In FIGURE 9, the two curves are not coincident. As expected, the ultracentrifuge curve spreads beyond the bounds of the diffusion curve. The sample is, therefore, extremely polydisperse. If one treats an ultracentrifugal curve as a normal diffusion curve, one would expect the apparent diffusion coefficient of a polydisperse system to increase both with time and with speed of rotation. Such was found to be the case for the methyl synthetic starch of FIGURE 9 and is confirmation of its polydispersity.[20]

---

[19] **Bevilacqua, E. M., E. B. Bevilacqua, M. M. Bender & J. W. Williams.** Ann. N. Y. Acad. Sci. **46** (5): 309. 1945.
[20] **Dunlap, R. I., Jr.** Dissertation. Columbia University. 1943.

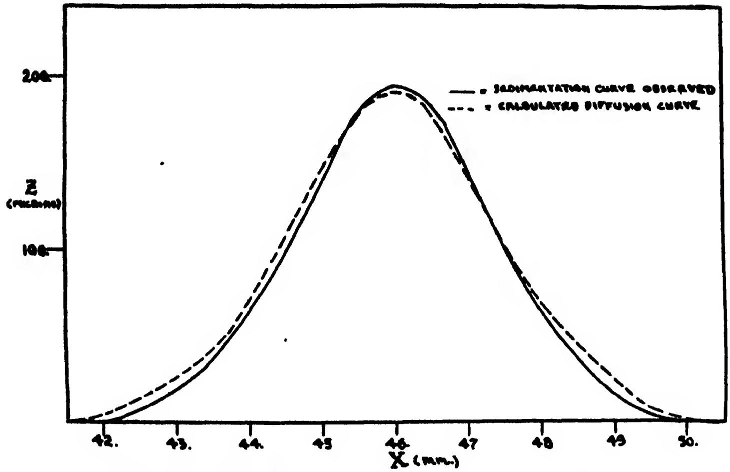


FIGURE 8. KAS corn amylose acetate in methyl acetate; t = 70 minutes

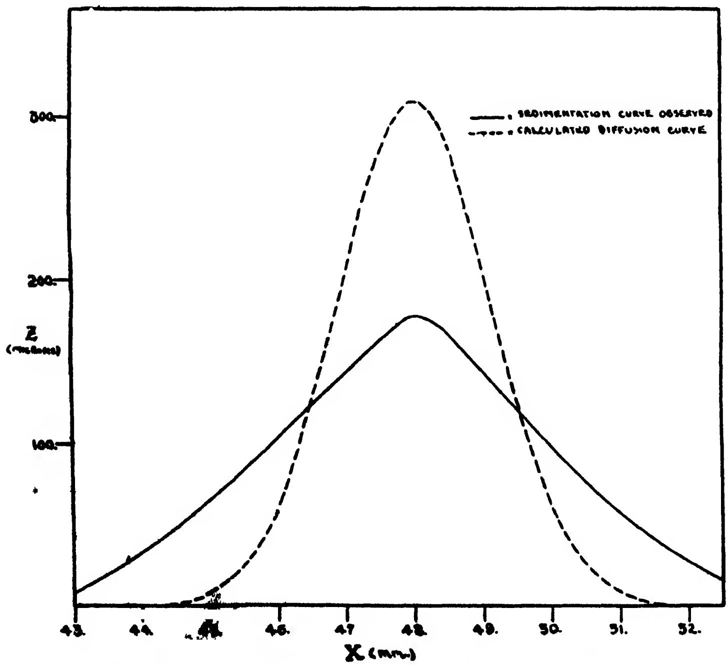


FIGURE 9. MSS methyl synthetic starch in methyl acetate; t = 40 minutes.

ACKNOWLEDGMENTS

The authors are indebted to Professor H Mark for placing at their disposal fractionated samples of cellulose acetate and polystyrene, together with some of his unpublished measurements on these samples.

This research was supported, in part, by the Corn Industries Research Foundation.

Annals of the New York Academy of Sciences
Volume XLVI, Art. 6. Pages 347–530
March 15, 1946

SURFACE ACTIVE AGENTS*

By

M. L. Anson, R. R. Ackley, Earl K. Fischer, David M. Gans, M. H. Hassialis, Rollin D. Hotchkiss, Donald Price, A. W. Ralston, Leo Shedlovsky, E. I. Valko

CONTENTS

* This series of papers is the result of a conference on Surface Active Agents held by the Section of Physics and Chemistry of The New York Academy of Sciences, January 26 and 27, 1945.
Publication made possible through a grant from the Conference Publications Revolving Fund.

INTRODUCTION TO THE CONFERENCE ON SURFACE ACTIVE AGENTS

By M. L. Anson

Continental Foods, Inc., Hoboken, New Jersey

The surface active agents which are the subject of this Conference are almost all water-soluble substances which, even in small concentration, lower the surface tension of water considerably. The typical structure is characterized by a large hydrophobic group and a hydrophilic group attached at some one point. The hydrophilic group may be positively or negatively charged, or may have no charge at all. The properties of the surface active agents depend primarily upon the size and shape of the hydrophobic group and on the charge and the location of the hydrophilic group. The exact chemical structure of the hydrophobic and hydrophilic groups, while of some importance, is of very much less importance than in ordinary chemistry. Since the properties of surface active agents depend primarily on very general considerations of geometry and charge, the reactions of surface active agents are enormously varied and relatively non-specific.

Some of the surface active agents have been known as chemical substances for a long time. It is only relatively recently, however, that surface active agents became available as cheap commercial compounds. In very short order, a bewildering array of surface active agents was prepared and a bewildering variety of important industrial applications was discovered. Most of the work on surface active agents has been done in industrial laboratories. The companies have been so much in a hurry to get new substances and new applications that the basic scientific work on pure substances has lagged and is only now beginning to be actively supported. It is our hope that this Conference will stimulate the whole tendency to put the understanding of surface active agents and their applications on a better scientific basis.

That there now exists such a great variety of surface active agents is partly due to the fact that surface active agents are today tailor-made for specific purposes. But there have also been two purely commercial reasons for the development of many surface active agents. First, companies have wanted to prepare surface active agents of types which they can patent and which have not been patented before. Secondly, different companies have interests in different raw materials. One

company will want to start with petroleum, another with fat, and a third, which has no normal basic connection with either petroleum or fats, will want to start with purely synthetic materials, preferably synthetic materials which it is already manufacturing.

We planned originally to have some very fancy discussions on the relation between the structure and the composition of surface agents and their properties and applications. This planning did not get very far before the participants in the Conference made it painfully clear to me that there is simply not enough knowledge of the properties of pure surface active agents to permit any great theoretical discussion of the relation between structure and properties. They will do the best they can and let it go at that.

The first day of this Conference deals with the properties of surface active agents and how they are measured, and how these properties are related to structure. The second day deals with the application of surface active agents to biology, medicine and industry. There is one kind of application to industry in which I am personally interested, which, unfortunately, has, as yet, not been carried very far. I refer to the application of surface active agents to the food industry. All sorts of very important possibilities exist, particularly in connection with dehydrated foods, which are now of such great military importance. A surface active agent to be useful in the food industry not only has to do a particular job in a particular food process, but it has to be non-toxic and almost tasteless.

Finally, I should like to mention two difficulties under which the speakers of this Conference have labored. The first difficulty, about which I have said something, is that the speakers are in the unpleasant position of trying to talk good science, when not enough good science exists. The second difficulty is that this Conference has been organized on much shorter notice than is usual for the Conferences of the Academy. This has made it necessary for the choice of speakers to be limited largely to men in this area. All of the speakers have helped in the planning of the Conference and I want to express my indebtedness to them.

THE STRUCTURE AND PROPERTIES OF SOLUTIONS OF COLLOIDAL ELECTROLYTES

By A. W. Ralston

From the Chemical Research Department, Armour and Company, Chicago, Illinois

The surface active agents offer an excellent subject for a symposium, not only because of the present day importance of surface active compounds, but principally because our theories concerning many of the fundamental principles which govern their behavior are still nebulous and, in many instances, highly controversial. Perhaps no group of synthetic chemicals goes back so far in history as do the soaps, and, certainly, none has been so universally used and yet so little understood. Until comparatively recently, one could complacently agree with the opinion of Krafft and others that soaps simply form colloids, whose peculiar properties were explained naively by many of the vague theories of classical colloid chemistry.

In any solution, we are dealing with several interdependent properties, namely, those of the surface and those in the main body of the liquid. When we consider surface active agents, such as the colloidal electrolytes, the dependence of the surface properties upon those of the main body of the solution is clearly evident. It is the purpose of this paper to discuss the structure and properties of solutions of colloidal electrolytes which pertain to the body of the solution itself, leaving for a future paper a consideration of those properties which relate primarily to the surface. It must, of course, be realized that, in solutions of colloidal electrolytes, there are actually many surfaces, such as colloid-solution or ionic micelle-solution interfaces, and that they probably do not differ fundamentally from those found at air-water interfaces.

Solutions of soaps or of other colloidal electrolytes are materially better conductors of electricity than would be predicted on the basis of their viscosity or osmotic effects. For example, a 0.2 N solution of potassium oleate has an equivalent conductivity of 33.3 mhos at 18° C. which, if due to potassium ions, would correspond to a concentration of .097 N, whereas the freezing point lowering corresponds to a concentration of only .065 N. There are, therefore, not sufficient simple ions present to account for the observed conductivity. In 1913, McBain[1] postulated that this high conductivity must be due to the presence of a

[1] McBain, J. W. Trans. Faraday Soc. 9: 99. 1913.

highly charged associated ion which he termed a micelle. The assumption that the high conductivity of soap solutions could be ascribed to hydrolysis was dispelled in the next year by the work of McBain and Martin[2] and later of McBain and Bolan[3] upon sodium and potassium soaps, of Goldschmidt and Weissmann[4] upon ammonium soap, and of Reychler[5] upon cetane sulfonic acid and cetyl trimethyl ammonium iodide which showed that hydrolysis in concentrated soap solutions amounts to only a fraction of 1 per cent. Thus, the high conductivities cannot be ascribed to a high concentration of free alkali, and the micelle theory received a substantial impetus, replacing the older concepts of soap solutions as simple colloids. In 1920, McBain and Salmon[6] justified the existence of ionic micelles on strictly mechanical grounds by the application of Stokes' law. This law may be expressed as $V = \frac{F}{6\pi r\delta}$ when applied to a sphere moving through a liquid of viscosity δ, the force, F, being due to the electric charge on the ion (96,540 coulombs per gram ion). These authors pointed out that if several ions, for example, twelve, associate, the driving force would be 12 F, whereas the radius of the sphere would be increased by only 2.3 times its original value. The calculated velocity would thus be increased to 5.2 times its initial value. This calculated velocity, however, would not be realized due to the hydration of the charged particle.

While the observation that the conductivities of solutions of colloidal electrolytes are higher than would be indicated from their osmotic and other properties is an interesting phenomenon, it is not until we study the slopes of their conductivity curves that we are faced with the fact that such solutions possess distinctive and characteristic properties differing fundamentally from those of simple ionic solutions. It can be safely stated that no single theory has thus far been advanced which adequately explains all of the observed phenomena. The work of McBain and Taylor[7] upon the conductivities of solutions of the sodium soaps; of Bunbury and Martin[8] upon the potassium soaps; of McBain and others,[9, 10] and of Wright, Abbott, Sivertz and Tartar[11] upon the sodium alkyl sulfates and sulfonates; and the recent work of Ralston,

---

[2] McBain, J. W., & M. E. Martin. J. Chem. Soc. 105: 957. 1914.
[3] McBain, J. W., & T. R. Bolam. Ibid. 113: 825. 1918.
[4] Goldschmidt, F., & L. Weissmann. Z. Chem. Ind. Kolloide 12: 18. 1913.
[5] Reychler, A. Bull. soc. chim. Belge. 27: 217. 1913.
[6] McBain, J. W., & C. S. Salmon. J. Am. Chem. Soc. 42: 426. 1920.
[7] McBain, J. W., & M. Taylor. Zeit. für phys. Chemie. 77: 179. 1911.
[8] Bunbury, H. M., & H. E. Martin. J. Chem. Soc. 105: 417. 1914.
[9] McBain, J. W., & W. D. Betz. J. Am. Chem. Soc. 57: 1905. 1935.
[10] McBain, M. E. L., W. B. Dye & S. A. Johnston. Ibid. 61: 3210. 1939.
[11] Wright, K. A., A. D. Abbott, J. Sivertz & H. V. Ibid. 61: 549. 1939.

Hoerr and Hoffman[12] upon the amine salts, show that the conductivity characteristics of all solutions of colloidal electrolytes are qualitatively similar and are comparable with one another. FIGURE 1 shows the

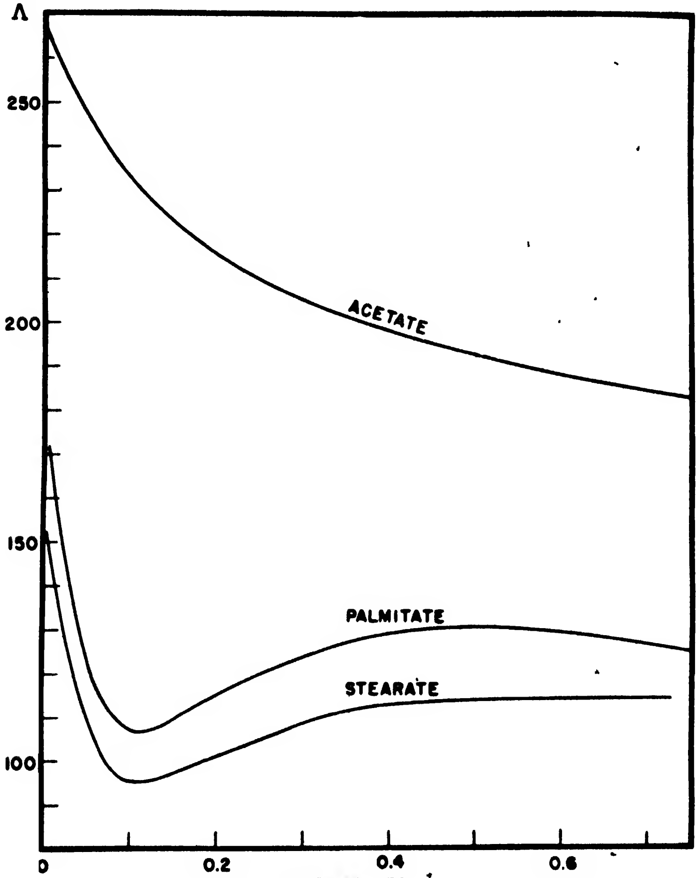


FIGURE 1. Equivalent conductivities of potassium acetate, palmitate and stearate (McBain, Laing and Titley).

equivalent conductivities of potassium acetate, potassium palmitate and potassium stearate plotted against molality, as published by McBain, Laing and Titley.[13] FIGURE 2 shows the equivalent conductivity

[12] Ralston, A. W., C. W. Hoerr & E. J. Hoffman. J. Am. Chem. Soc. 64: 97. 1942.
[13] McBain, J. W., M. E. Laing & A. F. Titley. J. Chem. Soc. 115: 1279. 1919.

of undecane, dodecane and tridecane sulfonic acids as compared to hydrochloric acid[9] and FIGURE 3 shows the equivalent conductivities of

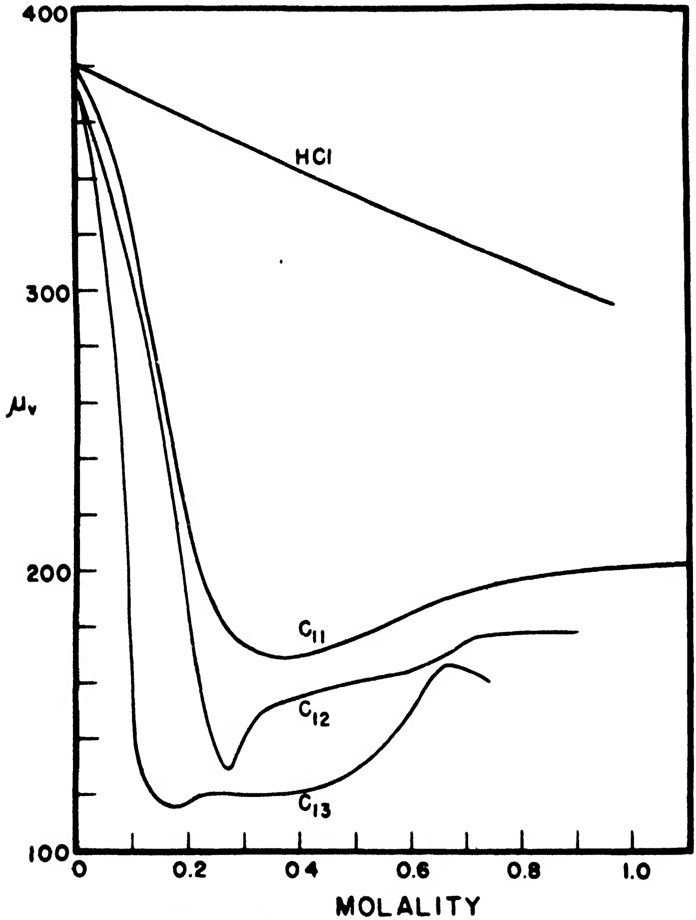


FIGURE 2. Equivalent conductivities of undecane, dodecane and tridecane sulfonic acids compared to hydrochloric acid (McBain and Betz).

solutions of typical cationic electrolytes, amine hydrochlorides, as determined by Ralston and Hoerr.[14] The similarity in appearance of the conductivity curves of these three types of colloidal electrolytes is

---

A. W., & C. W. Hoerr. J. Am. Chem. Soc. 64: 772. 1942.

quite striking. All show an initial slope characteristic of strong electrolytes and generally following, although not identical with, the theoretical Onsager slope. This portion of the curve where a normal behavior is observed is referred to as the first range. At a certain critical concentration (0.013 molar for dodecylamine hydrochloride and 0.0003 molar for octadecylamine hydrochloride), the equivalent conductivity

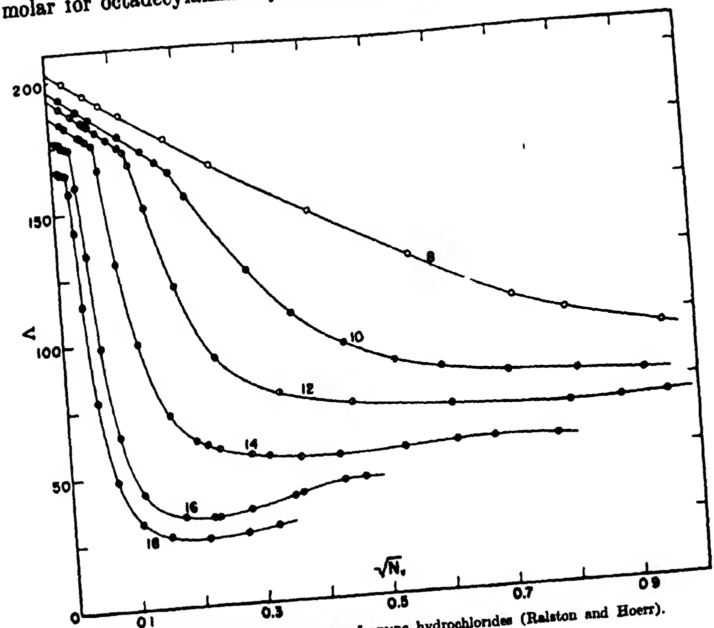


FIGURE 3. Equivalent conductivities of amine hydrochlorides (Ralston and Hoerr).

falls sharply, the drop occurring at a somewhat lower concentration with higher temperature. This is known as the second range and is typical of all colloidal electrolytes. The equivalent conductivity then rises or remains constant and this has been designated as the third range.

It will be noted that the slope within the first range is greater the higher the molecular weight of the amine salt and that the abruptness of the drop within the second range increases with increasing molecular weight. The rise in equivalent conductivity within the third range is more pronounced the higher the molecular weight. The effect of tem-

perature upon the conductivities of dodecylamine hydrochloride[12] is shown in FIGURE 4. An examination of these curves shows that, while different temperatures modify the slopes within the first and second ranges, the general forms of the curves remain qualitatively unchanged

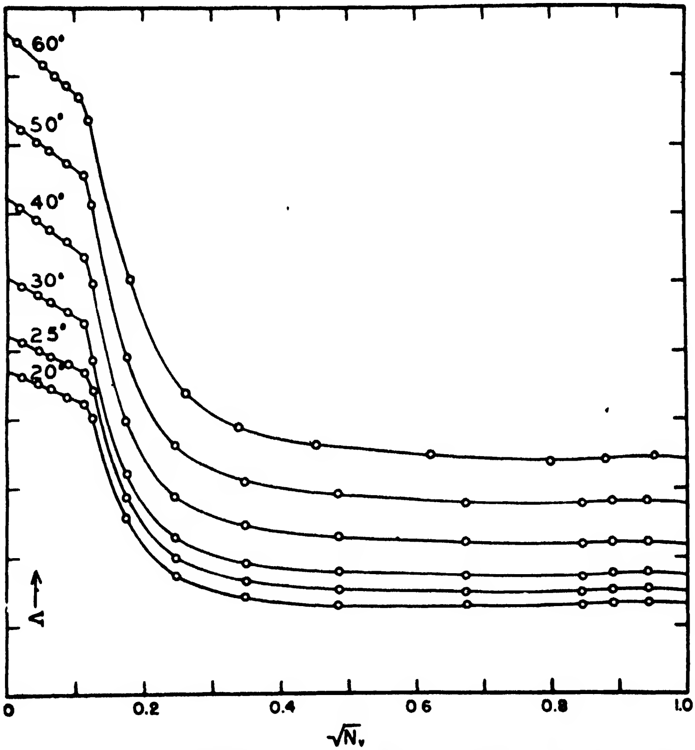


FIGURE 4. Equivalent conductivities of dodecylamine hydrochloride at various temperatures (Ralston, Hoerr and Hoffman).

over the temperature range shown. The higher the temperature the steeper is the slope within the first range and the lower the concentration at the critical point.

The sharp break in equivalent conductivity has been attributed by McBain and others to micelle formation. In order to explain this rapid fall these investigators have postulated that two types of colloidal particles are formed, the one a large "neutral" colloid contributing little or nothing to the conductivity and the other a highly charged ionic

micelle. These two types of particles are in equilibrium, the relative amounts of each being dependent upon the concentration of the solution and the temperature. The rise in equivalent conductivity in the concentrated solution is stated to be due to a shift in the equilibrium toward the formation of a greater proportion of the ionic micelle. It is apparent why the existence of two types of micelles is a logical assumption, since it is not reasonable that ionic micelles can first account for the drop in equivalent conductivity and later be responsible for the rather appreciable rise. Recent X-ray investigations[15-22] indicate that two forms of micelles differing in size and structure are actually present in solutions of colloidal electrolytes. One appears as a laminated particle consisting of alternate layers of undissociated molecules while the other is spherical and much smaller in size. Many years prior to these X-ray investigations McBain and Jenkins[23] had subjected soap solutions to ultrafiltration and had claimed a separation of ionic micelles of sodium oleate from neutral colloid, the latter apparently having a diameter in excess of 75 $\mu\mu$, while the former are much smaller. These results appeared to confirm the diagram previously developed by conductivity and freezing point measurements. This diagram is shown in FIGURE 5, and portrays the relative amounts of neutral colloid, ionic micelles, undissociated soap and simple ions plotted against the concentration of the solution. It will be noted that, in dilute solutions, we are dealing with acid soap, simple ions and simple molecules, together with some ionic micelles. In more concentrated solutions (0.05M), we have less acid soaps and simple ions and an increased amount of simple molecules. Increasing concentrations yield more ionic micelles, less simple ions and molecules and are accompanied by the formation of neutral colloid. At the higher concentrations, only ionic micelles and neutral colloid are present.

The results of sedimentation experiments which have been performed upon soap solutions[24] by use of the ultracentrifuge have been quite inconclusive. Svedberg[25] has attributed this observation to the fact that the partial specific volumes of the soaps in solution are quite close to unity, probably due to rather extensive hydration.

[15] **Thiessen, P. A., & E. Spychalski.** Z. physik. Chem. **156A**: 435. 1931.
[16] **Hess, K., & J. Gundermann.** Ber. **70B**: 1800. 1937.
[17] **Hess, K., W. Philippoff & H. Kiessig.** Kolloid Z. **88**: 40. 1939.
[18] **Stauff, J.** Ibid. **89**: 224, 1939; **96**: 244. 1941; Naturwissenschaften **27**: 213. 1939.
[19] **Kiessig, H., & W. Philippoff.** Ibid. **27**: 593. 1939.
[20] **Kiessig, H.** Kolloid Z. **96**: 252. 1941; **98**: 213. 1942.
[21] **Philippoff, W.** Ibid. **96**: 255. 1941.
[22] **Ross, S.** J. Phys. Chem. **46**: 414. 1942.
[23] **McBain, J. W., & W. J. Jenkins.** J. Chem. Soc. **121**: 2325. 1922.
[24] **McBain, J. W., & M. E. McBain.** Proc. Roy. Soc. London A **139**: 26. 1933.
[25] **Svedberg, T., & K. O. Pedersen.** The Ultracentrifuge. Clarendon Press. Oxford. 1940.

The contention that the behavior of solutions of colloidal electrolytes can be explained only by assuming the presence of two types of micelles

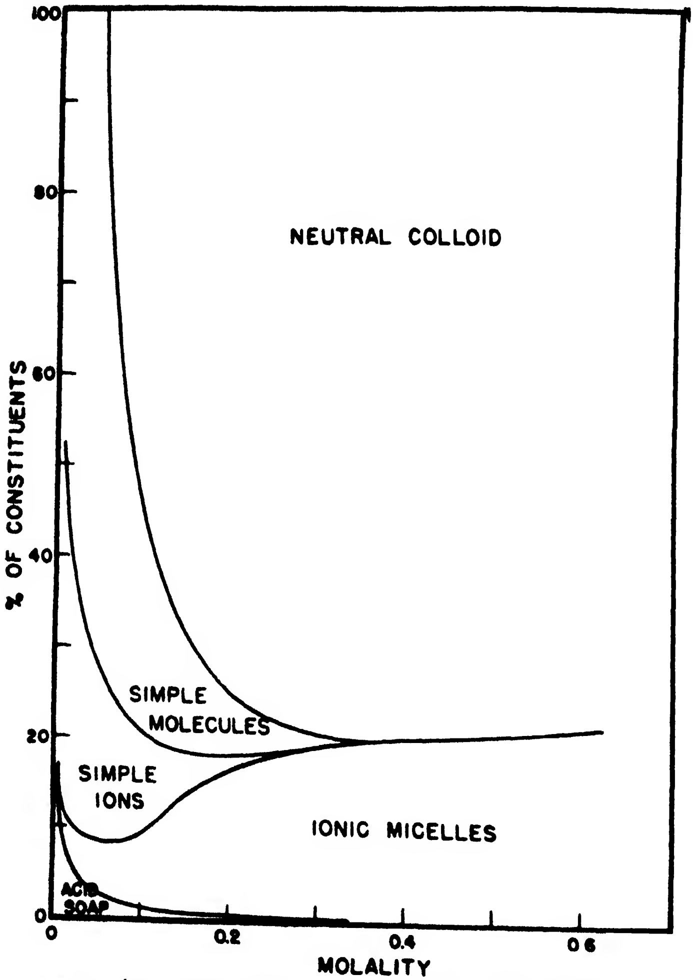


FIGURE 5. Physical state of sodium oleate solutions at various concentrations (McBain and Jenkins)

has been vigorously attacked by Hartley[26] on the basis that it disre-

[26] Hartley, G. S. Aqueous Solutions of Paraffin Chain Salts. Hermann et cie. Paris. 1936.

gards the effect of the various coulomb forces which exist between the ions and the micelles. Hartley believes that the observed properties of solutions of colloidal electrolytes can best be explained on the basis that only one type of micelle exists. This consists of an ionic micelle formed by the association of a number of high molecular weight ions, to which particle are attached a number of oppositely charged ions, termed

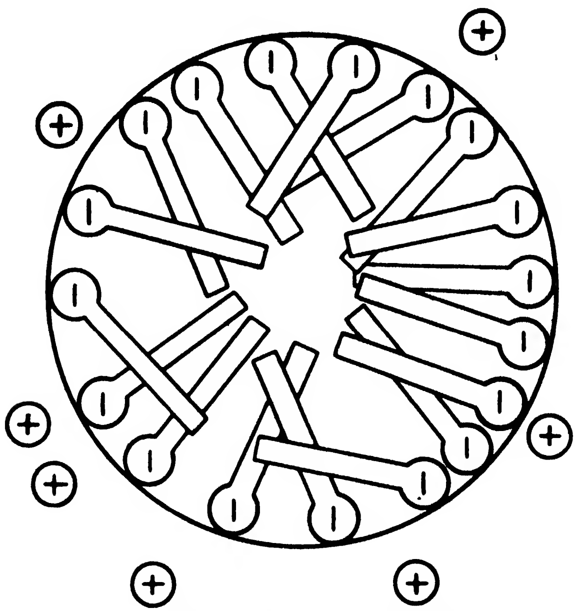


FIGURE 6. Structure of ionic micelle (Hartley).

"gegen-ions." The rise in equivalent conductivity of the concentrated solutions is attributed to an increased ionization of these gegen-ions from the ionic micelle. The structure of the micelle as visualized by Hartley is shown in FIGURE 6.

This discussion, so far, has purposely made no reference to conductivity effects of colloidal electrolyte solutions other than the relationship which exists between equivalent conductivities and concentration. The distinctive properties of this relationship have been shown to characterize solutions of colloidal electrolytes, and several theories based

upon the presence of micelles have been reviewed. When the transference numbers of the individual ions of colloidal electrolytes are considered, we encounter a phenomenon equally as characteristic and complex as that exhibited by the equivalent conductivities.

Migration data for a number of the colloidal electrolytes such as the potassium soaps,[27] lauryl sulfonic acid[28] and primary amine hydrochlorides[29] have been published. Detailed reference will be made to the migration data obtained for the amine hydrochlorides since they are quite typical of colloidal electrolytes in general. The amine hydrochlorides differ from the soaps, sulfonic acids, and alkyl sulfonates and sulfates in that the long hydrocarbon chain is in the positive or cationic portion of the molecule. They are representative of a large group of colloidal electrolytes known as cationic electrolytes, which comprise all the amine salts in addition to the many quaternary ammonium compounds. The cationic transference numbers of the amine hydrochlorides containing even numbers of carbon atoms from eight to eighteen, inclusive, are shown in FIGURE 7.

It will be noted that the transference numbers are in reasonable agreement with the expected values over that range of concentration where the conductivities of these salts showed that they function as simple electrolytes. The cationic transference numbers then rise abruptly at a point coincident with the break in the equivalent conductivities shown in FIGURE 3. The sharpness of this break in electrical properties has been commented upon by a number of investigators[11, 30-37] The longer the paraffin chain, the lower the concentration at which this rapid rise is evidenced Thus, it occurs at 0.032 molar for C_{10}, at 0.013 molar for C_{12} and at 0.0045 molar for C_{14}. The cationic transference numbers at infinite dilution decrease regularly with increased chain length. However, the transference numbers, after the break, increase in the order C_8, C_{10}, C_{18}, C_{16}, C_{12} and C_{14}. This is an interesting observation, since it means that whatever effects produce these abnormal transference numbers are at a maximum when the amine chain contains fourteen carbon atoms. In the third range, which was previously shown to be characterized by an increase in equivalent conductance, the

[27] McBain, J. W., & R. C. Bowden. J. Chem. Soc. 123: 2417. 1923.
[28] McBain, M. E. L. J. Phys. Chem. 47: 196. 1943.
[29] Hoerr, C. W., & A. W. Ralston. J. Am. Chem. Soc. 65: 976. 1943.
[30] Lottermoser, A., & F. Püschel. Kolloid Z. 63: 175. 1933.
[31] Bury, C. R., & G. A. Parry. J. Chem. Soc. 1935: 626. 1935.
[illegible] V. Tartar. J. Am. Chem. Soc. 58: 322. 1936.
[illegible] 347. 1936.
[illegible] Ibid. 61: 539. 1939.
[illegible] J. Phys. Chem. 43: 1173. 1939.
[illegible] J. Am. Chem. Soc. 61: 544. 1939.
[illegible] A. Piccione & C. Rosenblum. J. Phys. Chem. 46: 662. 1942.

cationic transference numbers decrease, the rate of fall being more rapid and occurring at lower concentrations with increased molecular weight.

Thus, it is apparent that the transference numbers, like the equivalent conductivity values, can be divided into three regions: the first, where the solution behaves as a simple electrolyte following generally the theoretical Onsager values for conductance and showing approximately the expected transference values; the second region characterized by an abrupt and rapid decrease in equivalent conductivities ac-

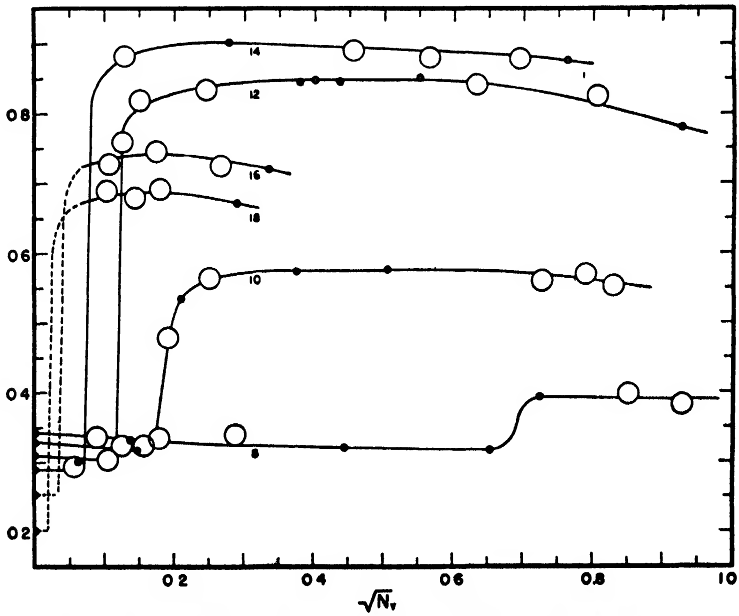
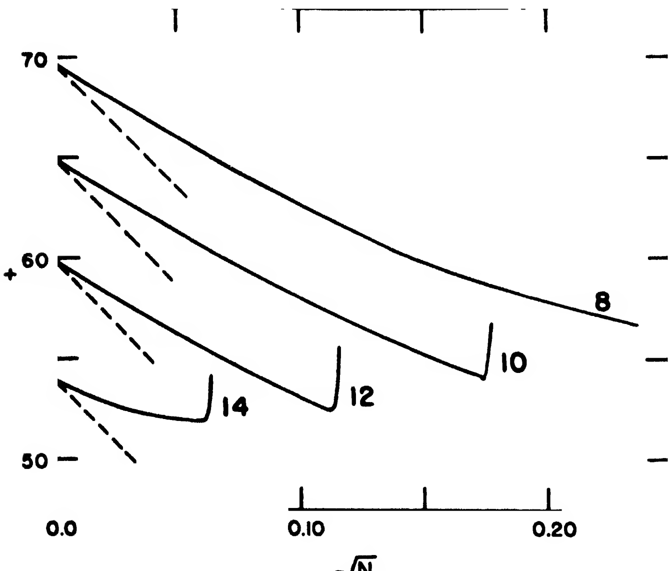


FIGURE 7 Cationic transference numbers of amine hydrochlorides (Hoerr and Ralston)

companied by an abnormal increase in the transference numbers of the cations (in the case of the amine salts); and the third region characterized by a slight decrease in the transference numbers of the cations and a somewhat gradual, but nevertheless definite, increase in the equivalent conductivities. Let us now consider these three regions particularly with regard to the other changes in physical properties which coincide with changes in the electrical behavior.

In the first range, it has been stated that these compounds function as simple electrolytes. This is, however, not strictly true, since the cationic conductances are substantially higher than the theoretical. FIG-

[illegible]ure 8 shows that the C_8 amine salt shows a departure from the theoretical conductance below the critical concentration, this deviation becoming decidedly more pronounced with increased molecular weight of the amine salt. Thus, it is apparent that micelle formation begins gradually, even in very dilute solutions, and it is possible that the sharpness of the break in electrical properties has been somewhat overstressed. The sharpness of this break has been often used as an argument against the micelle theory, since an abrupt change from a normal to an abnormal behavior cannot be justified by mass law considerations. Since Davies and Bury[38] have reported micelle formation in potassium octo-



[illegible] 8. Cationic transference numbers of amine hydrochlorides in dilute solution (Hoerr and Ralston).

ate solutions and Grindley and Bury,[39] in solutions of butyric acid, it is not surprising that octylamine hydrochloride shows some deviation from a normal behavior. The magnitude of the deviation of the cationic conductances when compared with those observed for the equivalent conductivities indicates that ionic micelles, rather than neutral colloid, are formed in this range.

[illegible], D. G., & C. R. Bury. J. Chem. Soc. 1930: 2263. 1930.
[illegible]y, J., & C. R. Bury. Ibid. 1929: 679. 1929; 1930: 1665. 1930

In the second range, we observe an abnomal rise in the transference numbers of the cation which is attended with a decided drop in equivalent conductivities. Increase in temperature is accompanied by increased cationic transference numbers for the dodecylammonium ion up to 40°C., after which lower values are observed. This region is characterized by a greatly increased solubility of the amine salt,[40] similar to that observed by Tartar and Wright[34] for the alkyl sulfonates. This change in the solubility of the alkyl sulfonates coincides with a break in the density and viscosity curves,[45] although a break in the viscosity curves has not been observed with the amine hydrochlorides in this range. McBain and others attribute this abrupt change in electrical and other physical properties to the rapid formation of lamellar and ionic micelles. The formation of a substantial portion of colloidal aggregates possessing low conductivities would account for the fall in equivalent conductivity, while the presence of an ionic micelle would explain the high cationic transference numbers It has been shown conclusively by McBain and Bowden[27] that the transference numbers observed cannot be explained on the basis of hydration. Hartley[26] maintains that only one type of micelle, the ionic micelle, is present and that the attachment of "gegen-ions" accounts for the lowered transference number of the anions. It does not appear that these two views are absolutely incompatible, since it is quite possible that some "gegen-ions" may be attached to the ionic micelle, thus accounting for lowered equivalent conductivity and high transference numbers in the case of cationic electrolytes. On the other hand, the attachment of a large number of "gegen-ions" to the micelles would so decrease their mobility, and consequently the conducting powers of the cations, that it is somewhat difficult to account for the extremely high transference numbers of the cations solely upon this basis. The greatly increased solubility within this range has been ascribed by Tartar and Wright[34] to the increased hydrophilic properties of the ionic micelle to which "gegen-ions" are attached. It appears, however, that increased solubility could be equally well explained on the basis of the formation of neutral colloid, since solubility would then not be confined to the limitations of an ionic system.

In the third range of concentration, the equivalent conductivities rise materially from the minimum values observed in the second range, and the transference numbers of the cations (for cationic electrolytes) decrease slightly, although abnormally high values are still obtained.

---

[40] Ralston, A. W., E. J. Hoffman, C. W. Hoerr & W. M. Selby. J. Am. Chem. Soc. 63: 1598. 1941.

McBain has ascribed this effect to a shift in the equilibrium between lamellar and ionic micelles toward a greater proportion of ionic micelle, while Hartley considers that it is due entirely to the liberation of "gegen-ions." Neither of these theories offers a completely satisfactory explanation. If the increased equivalent conductivity is occasioned by the formation of more highly conducting ionic micelles, one would expect an increase and not a decrease in the cationic transference values. The suggestion that these effects are due to the increased liberation of "gegen-ions" is, on the other hand, not completely acceptable, because of the obvious fact that dissociation tendencies are retarded by increased concentration. The explanation that these effects may be due to hydrolysis is equally unsatisfactory, Hoerr and Ralston[29] having shown that the rise in equivalent conductivity occurs at a point where the hydrogen ion concentration is not yet appreciable. In this third range, the solubilities of the amine hydrochlorides rise appreciably with increase of temperature and the viscosities suddenly increase to extremely high values. The dodecylamine acetate-water system[41] shows a gel area beyond 0.90 molar. There is, however, no appreciable change in the slope of the conductivity curve as the system goes from a solution to a gel. This increased viscosity can be attributed to the greatly increased concentration of lamellar micelles within this region.[21, 22, 42, 43]

Recently, Scott and Tartar[44] have obtained conductance data for aqueous solutions of bytyl-, hexyl-, octyl-, decyl- and dodecyltrimethylammonium bromide at 25°, 40° and 60°; and of hexadecyltrimethylammonium bromide at 25°. Octyltrimethylammonium bromide and its higher homologs showed downward breaks in their equivalent conductance curves at certain critical concentrations. When, on the other hand, the values of the equivalent conductivities of butyl- and hexyltrimethylammonium bromides were plotted against the square roots of their volume normalities, a linear relationship was observed. This indicates that a straight chain of at least eight carbon atoms is necessary for micelle formation in compounds of this type. Only hexadecyltrimethylammonium bromide showed a rise in equivalent conductance in the third concentration range. In a subsequent study[45] of two "double long chain salts," octyltrimethylammonium octanesulfonate and decyltrimethylammonium decanesulfonate, it was found that slightly conducting micelles form at concentrations much lower than for the corre-

Ralston, A. W., C. W. Hoerr & E. J. Hoffman. J. Am. Chem. Soc. 63: 2576. 1941; C. W. Hoerr & A. W. Ralston. Ibid. 64: 2824. 1942.

[42] McBain, J. W. J. Phys. Chem. 30: 239. 1926.

[43] McBain, J. W., M. J. Willavoys & H. Heighington. J. Chem. Soc. 1927: 2689. 1927.

[44] Scott, A. B., & H. V. Tartar. J. Am. Chem. Soc. 65: 692. 1943.

[45] Scott, A. B., H. V. Tartar & E. C. Lingafelter. Ibid. 65: 698. 1943.

sponding single long chain salts. A study[46] of the electrical properties of solutions of the sodium salts of straight chain alkyl benzene sulfonates indicated the benzene ring to be the equivalent of about three and one-half straight chain carbon atoms as regards its effect upon the critical concentration for micelle formation.

Recently, Van Rysselberghe[47] has attempted an explanation for the observed freezing point and conductivity data of aqueous solutions of lauryl sulfonic acid based upon the conception of a so-called "average micelle," the size and charge of which varies with concentration. The transference numbers below the maximum were shown to be in agreement with the calculated values.[48] While this concept may offer an explanation of some of the observed properties of colloidal electrolytes, the actual existence of an "average micelle" is incompatible with many of the properties of colloidal electrolyte solutions.

We have now reviewed the general properties of solutions of colloidal electrolytes, and discussed several of the more important theories which have been advanced to explain their behavior. It is of interest to speculate as to why micelles are formed and what forces may be involved in the determination of their size, osmotic effects and electrical properties. In this consideration, certain generalizations make themselves immediately apparent. Ionic micelle formation is essentially an electrical phenomenon and will be encountered only when substances capable of forming ions are involved. Likewise, we should not consider the lamellar micelle as simply a neutral colloid, but rather as a large particle possessing feeble ionic properties. Micelle formation will be encountered only in those systems whose components possess widely different internal pressures. If one dissolves oleic acid in acetonitrile, a certain critical concentration is reached, above which the system forms two immiscible liquids. Two phases will be encountered over an appreciable range of temperature and concentration. This simply means that, under these conditions, the attraction of the long alkyl chains for each other is greater than that for the solvent molecules so that a liquid phase high in acid and low in acetonitrile separates. No micelle formation is evidenced in this system. When, on the other hand, one dissolves octadecylamine hydrochloride in water, a certain critical concentration will be reached at which point the solubility, electrical and other properties undergo an abrupt change. This point, certainly, has something in common with that observed in the

[46] Scott, A. B., & H. J. Tartar. J. Am. Chem. Soc. 65: 692.
[47] Van Rysselberghe, P. J. Phys. Chem. 43: 1049. 1939.
[48] Van Rysselberghe, P. Ibid. 48: 62. 1944.

oleic acid—acetonitrile system, since it is the one at which the concentration of long alkyl chains is such that their attraction for each other becomes an important factor in determining the properties of the system. Because of the highly polar nature of the terminal groups and of the solvent, surface active properties are appreciable. The tendency of the head groups to orient towards the solution phase tends to produce a system having a maximum surface. Thus, instead of separating into two distinct liquid phases, the system separates into a liquid phase containing a large number of molecular aggregates. The particles can contain many molecules and are so constituted that the polar groups tend to be oriented toward the aqueous phase. Such aggregates may exhibit some ionic effects owing to a small amount of ionization of the head groups. Their properties, however, generally follow quite closely those of a colloidal particle. It is certain that a major distinction between the oleic acid—acetonitrile system and the octadecylamine hydrochloride-water system is that, in the latter, the surfaces are tremendously greater. This type of solution is probably encountered in the amine-water systems,[49] but, since the original deduction of the phase rule excludes the influence of surface and boundary effects, it is not apparent in the ordinary phase rule diagrams. The accentuation of the difficulty of defining the homogeneity of a phase, where colloidal systems are concerned, has been pointed out.[50] The formation of the lamellar micelles accounts for the rapid fall in equivalent conductance. Since they are possessed of rather feeble electrical properties, due to their low ionic properties and mobilities, their relative size is unimportant, within limits, in determining the electrical properties of the system as a whole. One important consideration with reference to these particles is that they must be in equilibrium with their surroundings, both as regards surface and ionic forces.

Simultaneously with the formation of these lamellar micelles, there will be formed a small number of ionic micelles in equilibrium with the larger particles. There is evidently a natural tendency for individual long chain ions to associate and form ionic micelles, as evidenced by the fact that ionic micelle formation occurs even in very dilute solutions. The formation of ionic micelles at the critical point is, therefore, not a spontaneous effect, as shown by the transference numbers in the dilute solutions. What really happens at the critical point is that the solubility of the amine hydrochloride is exceeded and the major characteristics of the system are, therefore, modified. It is evident that, before the

[49] Ralston, A. W., C. W. Hoerr & E. J. Hoffman. J. Am. Chem. Soc. 64: 1516. 1942.
[50] McBain, J. W., R. D. Vold & M. J. Vold. Ibid. 60: 1866. 1938.

critical point, the ionic micelle exists as an associated ion. The electrical properties show that, beyond the critical point, where the ionic solubilities are exceeded, another type of particle appears. The rise in transference numbers can be explained on the basis that the presence of lamellar micelles favors the formation of ionic micelles or that the particle which separates owes its feeble electrical properties essentially to "gegen-ion" attachment. A study of the correlation which exists between solubility effects and the attendant change in other physical properties will undoubtedly help to clarify this question.

The ionic micelles are approximately spherical in shape although they may be somewhat distorted due to their motion. Let us assume them to be spheres, and that their diameter is approximately twice the molecular length. The packing of the molecules within this sphere is apparently determined by several factors, such as the cross-sectional areas of the hydrocarbon chains and the area occupied by the head groups. The individual molecule is now oriented within this particle both by the van der Waals attractive forces of the chains and by the repulsions of the ionic groups in the surface layer. Since the surface of a sphere decreases rapidly with decrease in diameter, it is evident that this particle would not be stable were it not for the repulsion of the head ions. The size of the micelle will, therefore, be limited either by the cross sectional areas of the hydrocarbon chains or by the fields of repulsion exhibited by the like ionic group. It would thus appear that ionic micelles produced by the association of like ions should be of uniform size. Recently, Smith and Pickles[51] have shown that the micelles formed by digetonin are of approximately uniform size.

While some distortion may be produced by micelle formation between ions of different hydrocarbon chain lengths, the magnitude of this effect is problematical. Experiments which have attempted to show that the size of ionic micelles varies over a considerable range could be explained by the fact that compounds of different chain lengths were present initially. The concept of the ionic micelle structure which has been suggested is open to the criticism that an excessive degree of hydration would be encountered, because of the fact that the chains are separated much further as they approach the surface of the sphere. The internal structure of the micelle, being composed of hydrocarbon chains, would, however, be antagonistic to a high degree of hydration. It appears quite possible that both the cross sectional areas of the hydrocarbon chains and of the head groups, combined with the ionic repulsion between like ions, function to determine the size of ionic micelles, and

[51] Smith, E. L., & E. G. Pickles. Proc. Nat. Acad. Sci. 26: 272. 1940.

it is certainly evident that this particle would not be stable except for ionic forces.

DISCUSSION OF THE PAPER

Dr. E. I. Valko (*Onyx Oil and Chemical Co., Jersey City, N. J.*):

Discussing Hartley's conception in relation to the second range of concentration in which the abnormal rise in the transference number of the surface active ion coincides with the drop in equivalent conductivity, Dr. Ralston points out that "the attachment of a large number of gegen-ions to the micelles would so decrease their mobility, and consequently the conducting powers of the cation, that it is somewhat difficult to account for the extremely high transference numbers of the cations solely upon this basis." However, it seems to me that Dr. Ralston's paper lists the facts which account for this paradoxical behavior solely upon the basis of the attachment of the gegen-ions. Suppose the micelle is formed by association of thirty surface active ions. In this case, the electrical driving force is increased thirty-fold, but the radius only 3.3-fold, and, according to Stokes' law, the calculated mobility would be increased to nine times its original value. Experimentally, the increase of mobility is only three-fold, which means that two-thirds of the gegen-ions may be attached to the micelles and the McBain-effect would still account for the increased mobility of the surface active ions. The attachment of two-thirds of the gegen-ions would, of course, fully explain the extremely high transference number of the surface active ions.

I would like to mention here three independent confirmations of the size of the spherical micelles deduced by Hartley on the basis of geometrical considerations. These are: first, diffusion measurements by Hartley and Runnicles[1] on cetyl pyridinium chloride and cetyl sulfonate; second, ultracentrifuge measurements by Miller and Andersson[2] on sodium dodecyl sulfate; and, third, diffusion measurement by Hakala[3] on sodium dodecyl sulfate.

Dr. Ralston* emphasized the possibility of participation of ionic micelles in the biological and, particularly, the bactericidal action of surface active ions. We have recently had occasion to check this possibility by a few calculations and experiments. Dodecyl amine hydrochloride kills *Staphylococcus aureus* and *Eberthella typhosa* in about 1:10,000 dilution in five minutes. This concentration (approx. 5×10^{-4} M) is much lower than the critical concentration at which, according to the measurements of Ralston and Hoerr,[4] dodecyl amine hydrochloride begins to exhibit the abnormal conductivity behavior characteristic for the micelle formation. However, it must be taken into consideration that the germicidal activity by the F.D.A. method for the determination of phenol coefficients was not measured in distilled water, but in the presence of electrolytes introduced with the nutrient medium. Since the presence of electrolytes is likely to promote the micelle formation, we determined the germicidal effect of dodecyl hydrochloride, also in distilled water, using washed suspensions of the bacteria and found that, under these conditions, the killing concentration was substantially higher than in the presence of electrolytes, but still definitely lower than the critical concentration. The possibility, stressed by Dr. Ralston, that some micelles are present below the critical concentration, is admitted. Since, however, the forces acting between the surface active ions and the biological substrates, especially the proteins, are stronger than

[1] [illegible]y, G. S., & B. F. [illegible] Proc. Roy. Soc. London A 934: 420. 1938.
[2] [illegible], G. L., & E. J. I. A[illegible] J. Biol. Chem. 144: 475. 1942.
[3] [illegible], N. V. Dissertat[illegible] University of Wisconsin 1943, quoted by Bevilacqua,
[illegible] Bender. Ann. N. Y. Acad. Sci. 46 (5): 326. 1945.
[4] [illegible] Hoerr. J. Am. Chem. Soc. 64: 772. 1942.

the forces acting in the micelles between the surface active ions themselves, it is likely that, in reacting with these substrates the micelles dissociate into single ions. Due to the higher mobility and higher surface activity of the single ions, it is not unlikely that even in solutions which contain the major fraction of the surface active ions in the form of micelles, the single ions representing the minor fraction are primarily responsible for the biological effects. Since there is a dynamic equilibrium between micelles and single ions, adsorption of the single ions will cause dissociation of the micelles.

SURFACE ACTIVE AGENTS AT INTERFACES

By Earl K. Fischer* and David M. Gans†

Interchemical Corporation Research Laboratories, New York, N. Y.

In this paper, we propose to examine the rational foundation for the use of surface active agents in so far as it can be logically correlated with the facts of surface chemistry. This requires (a) a general view of properties measured at the air-liquid, liquid-liquid, and solid-liquid interfaces by a variety of experimental methods; (b) an evaluation of factors which affect the validity of such measurements and their interpretation, many of which have been given only scant attention; and (c) a brief survey of important industrial processes and products, where the application of quantitative methods assists, not only in understanding the properties observed, but also aids in planning research and development.[1]

As examples of the application of the concepts of surface chemistry, there may be mentioned the following: foam formation, which, in its simplest form, represents primarily the air-liquid interface; emulsification, which is dependent chiefly on relations at the liquid-liquid interface; dispersion of solids in liquids, which presents the important and complicated case typified by numerous commercial products. Detergency requires a consideration of all three of these main classes of interfaces. It is apparent that no single physical measurement will suffice to define all observed phenomena in these instances, although such an effort has frequently been made.

Our point of view seeks a compromise between the imperious demands of the moment and the slow, sure progress to be achieved with sound theory and fundamental data. There is, of course, no choice for the technician supervising factory operations. He must adapt some quick laboratory test, simulating in all details the factory operation to be controlled for output rate and for quality standard. But such tests, regardless of the rationale of their use, cannot be translated into principles, nor is it possible to project the interpretation of such data into other operations.

Basic to all the phenomena of the interface is the energy change involved. Ordinarily, a surface active agent is useful because it effects a

---

* Present address: Institute of Textile Technology, Charlottesville, Virginia.

† Present address: Quaker Chemical Products Corporation, Conshohocken, Pennsylvania.

[1] Fischer, E. K. Soap and sanitary chemicals 19 (12): 25–29, 53. 1943; 20 (1): 28–31, 67–69. 1944.

decrease in free energy at the interface. Many of the experimental techniques devised are intended to measure, directly or indirectly, the magnitude of this change. The easiest procedures are naturally the ones most commonly used.

THE AIR-LIQUID INTERFACE

Surface-tension data, as determined by various methods, have been reported for many commercial species of surface-active agents. Reviews on the general subject of surface chemistry have been given recently.[2] The range of values shown in FIGURE 1 represents the general region and does not, of course, indicate the complicated relationship between concentration and surface tension, especially at dilute concentrations (below 0.1% or ca. 0.002M.).

Surface-Tension Minima. In the low-concentration range, a dip in the curve of surface tension-concentration is often observed (FIGURE 2). This phenomenon has been known for some time,[3] but only recently has much attention been given to the interpretation, still incomplete, of a large mass of experimental data which bears on this point.

It should be pointed out that, although the surface tension versus concentration curve seems smooth, nevertheless, it harbors a rapid change in one concentration region. This is brought out by the Gibbs adsorption equation

$$\Gamma = \frac{-1}{RT} \frac{d\gamma}{d \ln c}$$

where Γ is the Gibbs surface excess of solute in moles per cm^2, c is the concentration (properly the activity), γ the surface tension, R the gas constant and T the absolute temperature. Now γ plotted against in c is a knee-shaped curve. For extremely dilute solutions where γ is only a dyne or two below that of pure water, Γ is very low and changes little. At higher concentrations, where γ assumes its lowest values, Γ reaches its maximum and approximately constant value. However, for solutions of intermediate concentration, for which γ is perhaps 5 to 10 dynes lower than for water, Γ changes rapidly from a very low to its maximum value. This narrow range in c for which $\frac{d\Gamma}{dc}$ shows a very

[2] ——ler, J. Colloid Chemistry 5. Reinhold Publishing Co. New York, N. Y. 1944. See sections as follows: Harkins, W. D.: 12–102; McBain, J. W.: 102–120; Vold, R. D., & M. J. Vold: 209–280. Cf. also Advances in Colloid Science. Chapters by McBain, J. W.: 99–142 and Hauser, E. A.: 391–415. Interscience Publishers. New York, N. Y. 1942.

[3] Harkins, W. D., R. E. Davies, & G. L. Clark. J. Am. Chem. Soc. 39: 541–596. 1917.

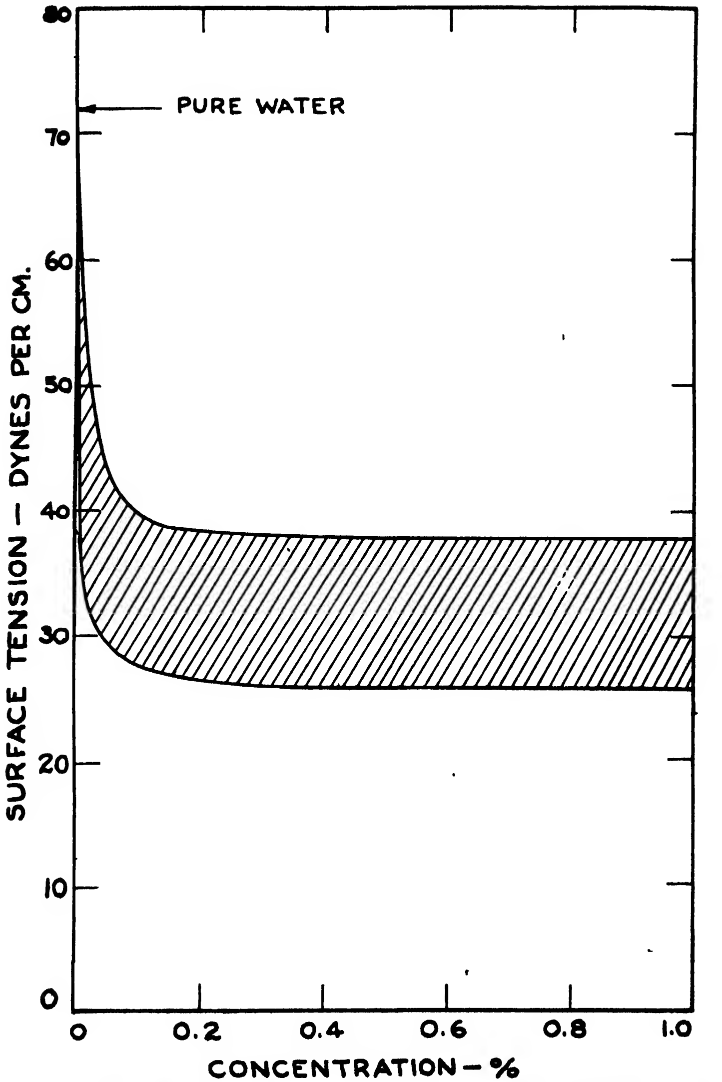


FIGURE 1A. Surface-tension values for majority of surface-active agents fall within shaded area on graph.

rapid rise, comes in or near the region exhibiting other anomalies, and may be one of the contributing factors, particularly where two compet-

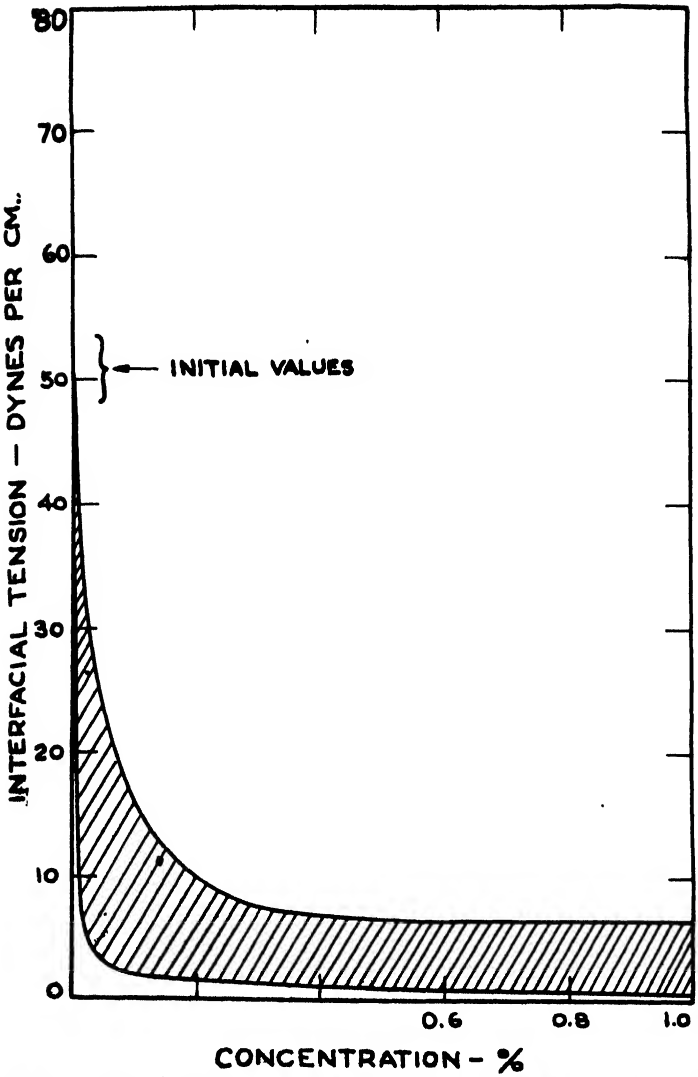


FIGURE 1B. Interfacial-tension values between mineral oil and water for the majority of surface-active agents fall within shaded area.

ing adsorbable species (one perhaps an impurity) are present, achieving their maximum adsorption possibly at different concentrations.

An abrupt dip in the surface tension curve, such as that often experimentally determined, showing a change in slope from negative to positive (implying negative adsorption) beginning at b (FIGURE 2), has been interpreted as a failure of the Gibbs adsorption isotherm, which has, however, successfully weathered many another and probably also

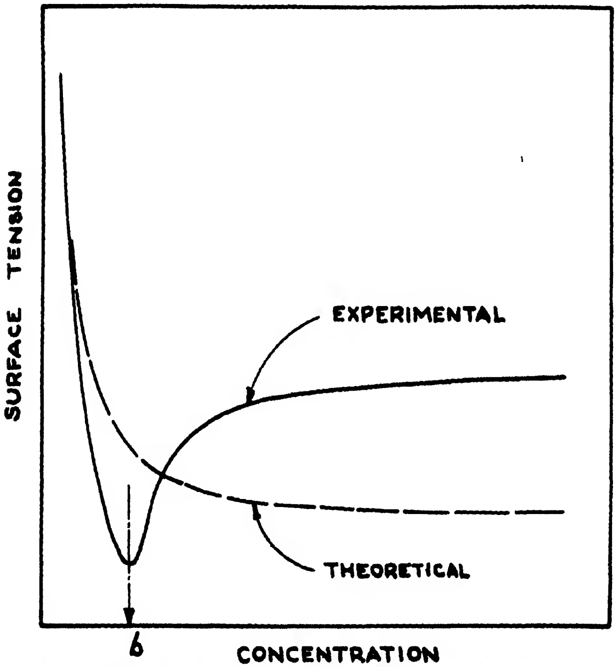


FIGURE 2 Generalized curves showing experimental relation, frequently observed, and the theoretical relation between surface tension and concentration of surface-active substances in aqueous solutions.

this criticism. There is evidence that the dip is to be explained otherwise.

The position of the dip is known to be affected by many factors. Electrolytes shift b to lower concentrations. Powney and Addison[4] observed that the "critical concentration" was shifted to lower values as the length of the hydrocarbon part of the molecule in paraffin-chain salts was increased. For example, the critical concentration for C_{12}

[4] Powney, J., & C. C. Addison. Trans. Faraday Soc. 33: 1243–1260. 1937.

compounds was 0.008 molar while that of the corresponding C_{18} compound was 0.00017. The values given are based on surface-tension measurements at 60° C. Lower temperatures resulted in a lower critical concentration. Powney and Addison noted a correspondence of values for critical concentration deduced from surface tension, interfacial tension, and conductivity data.

Experimental data have been presented by Miles and Shedlovsky[5] which show that surface-active compounds (fatty alcohol sulfates) in a high state of purity do not exhibit the characteristic minima, but that the addition of trace quantities of electrolytes or fatty alcohols produce minima. In the experiments of Bulkeley and Bitner[6] no minima were noted with sodium oleate solutions when precautions were taken to exclude carbon dioxide from the system during measurement. This evidence is in accord with the view[7] that, if additional surface-active species are present, either as single surface-active molecules or as aggregates in the form of micelles, complications ensue. Alexander[8] points out that, at *b* where $d\gamma/dc$ becomes zero, the concentration or activity of the molecular surface-active species may become almost independent of the stoichiometric concentration. The micellar components may be considered as reservoirs (or buffers) supplying molecular surface-active species at a rate which keeps the concentration of the latter nearly constant. This explains, however, only a flattening but not a dip in the surface tension data.

It is in this low-concentration region that other apparently anomalous effects are observed. Of these, the time necessary to reach an equilibrium condition of surface activity is of extreme importance, and we shall refer to this phenomenon again.

Attainment of Equilibrium. The time factor, in particular, has attracted attention from the earlier days of surface physics. In a spectacular experiment, Rayleigh[9] demonstrated the relatively slow attainment of the surface tension of a soap solution. Using the vibrating-jet method, in which the surface is formed and measured within 0.01 second, Rayleigh found that the surface tension of a 2.5% sodium oleate solution was close to that of pure water (ca. 73 dynes per cm.), although the capillary-rise method on the same solution gave a value of 25 dynes per cm.

---

[5] **Miles, G. D., & L. Shedlovsky.** J. Phys. Chem. **48**: 57–62. 1944.
[6] **Bulkeley, R., & F. G. Bitner.** Bur. of Standards Jour. of Research. **5**: 951–956. 1930.
[7] **Adam, N. K.** Physics and Chemistry of Surfaces. Third Edition. The Clarendon Press. Oxford. 1941.
[8] **Alexander, A. E.** Trans. Farad. Soc. **38**: 54–63. 1942.
[9] **Rayleigh, (Lord).** Nature **41**: 566–568. 1890.

This delay in reaching equilibrium has since been studied by a number of investigators and some of the anomalies in surface-tension data are in considerable measure attributable to the time factor alone.

With pure liquids, the surface tension is attained very quickly ($<$ 0.003 second), but in solutions the time for diffusion of the solute molecules into the surface and the reorientation in the surface may be very long—a matter of weeks for extremely dilute solutions. Accurate surface tension measurements in this range are rare. It has been observed that the fall in surface tension for sodium cetyl sulfate solutions, rapid at first, continues at a rate linear with time for many hours thereafter.[10] These results were for solutions 0.00001 to 0.01 N. in concentration. At higher concentrations, the surface tension reaches an equilibrium value more quickly. Added electrolytes hasten the attainment of steady readings. Adam and Shute[11] noted that the very dilute solutions exhibited decreasing surface-tension values, but that, ultimately—a week or longer, the values approached those of more concentrated solutions, which were attained quickly. A critical concentration (around 0.001 N. for hydrocarbon chain lengths of 16 carbon atoms and 0.005 N. for 12 carbon-length compounds) was found, above which the equilibrium concentration was reachcd very quickly. It has been suggested that the critical concentrations are those at which ionic micelles form. Other investigators have observed similar effects for a large variety of specially purified compounds, but find that several hours aging is sufficient for practical purposes and that, in the concentrations corresponding to the flat part of the curve, the time effect is negligible.[12]

Other phenomena may be cited to illustrate the time effect on surface activity. Adam[7] notes that hysteresis ("elastic after-working") of the unimolecular films of hydrolecithin and dodecyl phenol had been observed. One of the authors (F) observed a distinct lag between the compression and expansion values for myristic acid in the region of phase transition, both for film pressure and surface potential measurements.

Adsorption at the air-liquid interface which accompanies surface-tension changes has been studied directly by McBain and his associates[13] and by Gans and Harkins[14] (FIGURE 3).

G. C., F. A. Long & W. D. Harkins. J. Am. Chem. Soc. **62:** 1496–1504. 1940.

[11] **Adam, N. K., & H. L. Shute.** Trans. Farad. Soc. **34:** 758. 1938.

[12] **Dreger, E. E., G. I. Keim, G. D. Miles, L. Shedlovsky & J. Ross.** Ind. Eng Chem. **36:** 610–617. 1944; **Miles, G. D., & L. Shedlovsky.** J. Phys. Chem. **48:** 57. 1944.

[13] **McBain, J. W., & G. P. Davies.** J. Am. Chem. Soc. **49:** 2230. 1927; **McBain, J. W., & R. DuBois.** Ibid. **51:** 3534. 1929.

[14] **Gans, D. M., & W. D. Harkins.** J. Phys. Chem. **35:** 722–739. 1931.

In general, the amount of solute exceeded that computed for a unimolecular film. An explanation of the results required introduction of such factors as rate of adsorption and accumulation of solute at a surface of fluctuating contour.

It is of interest for our purposes, however, to note that, in those intermediate concentrations most commonly employed industrially, many of the anomalies shown by solutions of surface active agents diminish in importance, although it is necessary to keep in mind the fact that, under certain manufacturing conditions, a process depending on sur-

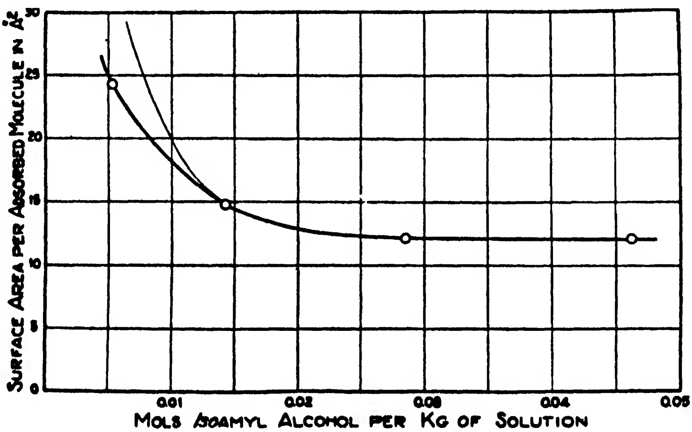


FIGURE 3. Variation of molecular area with concentration of solution

face-active properties may fall into the region where prediction becomes uncertain. An illustration of this effect, the exhaustion of a solution by continued processing, will be noted later.

LIQUID-LIQUID INTERFACE

The range of values for the interfacial tension between aqueous solutions of commercial surface active compounds and mineral oil (FIGURE 1) are average values determined by several methods. It will be noted that extremely low values may be reached, with the majority in the range of 1 to 3 dynes per cm. An unusually low value of 0.04 dyne per cm. for an aqueous sodium oleate-benzene system has been reported.[15]

W. D., & H. Zollman. J. Am. Chem. Soc. 48: 69–80. 1926.

The shape of the interfacial tension-concentration curve for most surface active substances bears a close resemblance to that of the analogous surface-tension curve. There is a rapid drop to a low value, followed by a rise and then a continued decrease. This "break" in the curve occurs between 0.05 and 0.1% for a sodium alkyl sulfate reported by Aickin[16] and the position of the break is independent of the oil phase. Addition of electrolytes shifted the critical concentration and effected a marked lowering of interfacial tension, a result which was found to be due almost exclusively to that of the ions of sign opposite to that of the surface active ion. These anomalies have been explained on the basis of a change with concentration in the ratio of the several possible ionic or molecular species.

An interfacial tension value of 10 dynes per cm. or less is sometimes given as a rough limit for ease of emulsification, and, at a value of 1 dyne per cm., many systems appear to emulsify spontaneously. The following observation may be made in the course of interfacial tension measurements on oil-water systems: if the two liquids are allowed to stand, the interface, at first, clear, becomes clouded to a depth of a millimeter or more. Spontaneous emulsification of this nature has been found to predict long emulsion stability as well as ease of emulsification at phase-volume ratios which preclude the formation of dual or revertible emulsions.

Interfacial Films. The emulsifying film at the liquid-liquid interface has been studied by a direct technique.[17] The specific interfacial areas of the dispersed oil globules in emulsions were measured by direct microscopy. From analytical data on the change in concentration of the emulsifying agent, it was possible to compute the area per molecule in the adsorbed film. Thus, it was found that the initial area for the sodium oleate molecule was of the order of 45 sq. Å, and that the area decreased gradually along a smooth curve to a terminal equilibrium value of about 20 sq. Å (FIGURES 4 and 5). This result was obtained on emulsions in which the initial sodium oleate concentration was 0.2 Molar. At higher soap concentrations, a condensed monomolecular film is obtained very rapidly and the emulsions so formed are stable (shelf storage) for periods as long as 10 years.

Emulsions formed from dilute soap solutions are stable for somewhat shorter periods, and it is rather remarkable that the interface is populated with one molecule for about 50 sq. Å or more of space. On

---

[16] **Aickin, R. G.** J. Soc. Dyers and Colourists **60:** 36–43 1944 Additional data, particularly with reference to detergency, are given by Aickin in succeeding papers, viz., J. Soc. Dyers and Colourists **60:** 60–65; 170–176, 286–287 1944

[17] **Fischer, E. K., & W. D. Harkins.** J. Phys. Chem. **36:** 98–110. 1932

aging, however, the specific interfacial surface decreases by coalescence of particles, facilitated by creaming of the dispersed oil globules, which, combined with the adsorption of an additional emulsifying agent from the solution, results in a film closely packed with soap molecules. At this

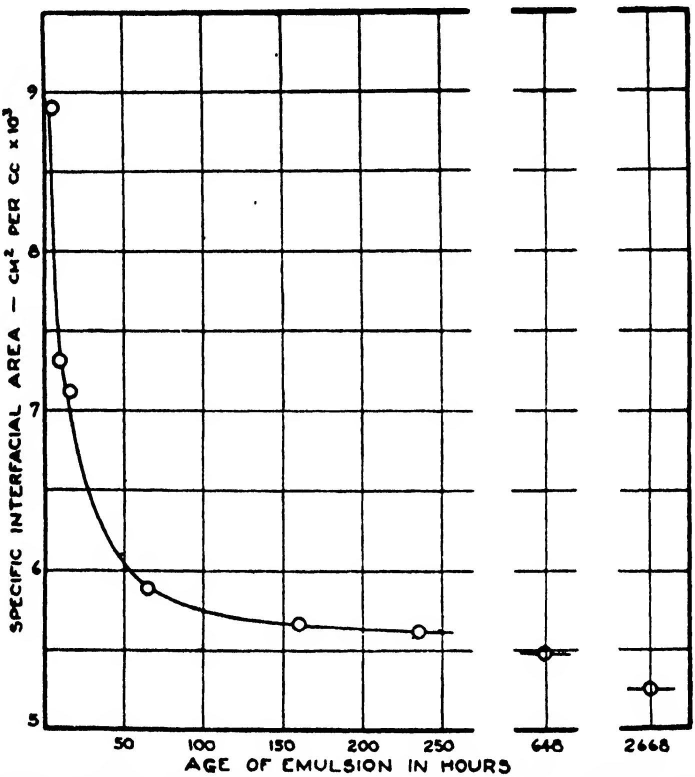


FIGURE 4. Decrease in interfacial area with the age of an emulsion in which the emulsifying agent was 0.02 molar sodium oleate.

stage, there is, presumably, sufficient lateral compression and rigidity in the film to prevent further coalescence.

Aging of the emulsion can be followed with certainty only by particle-size distribution counts on the actual emulsion at increasing time intervals. Although this is a tedious method, unattractive to most investigators, it has shown that the progressive decrease in interfacial area is the primary factor in evaluating emulsion stability. Examina-

tion of the gross features of the emulsion is not sufficiently sensitive to establish stability relations with finality.

A direct measurement of the interfacial film by means of the film-balance technique was made by Askew and Danielli[18] using bromo-benzene and water as the liquid pair. The float was at the interface of the two liquids. Many substances spread at this interface in a manner

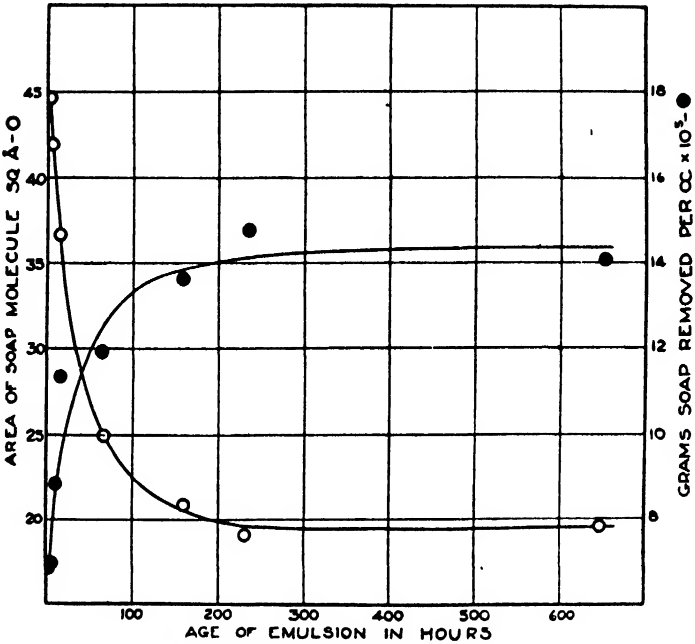


FIGURE 5. Decrease of the molecular area of sodium oleate in emulsion with its age

directly analogous to that of the air-liquid interface, although the numerical values were, as would be expected, somewhat different in magnitude.

Further information on the interfacial film was obtained by Alexander and Teorell[19] by means of a modified ring method which allowed a simultaneous measurement of the film pressure. At low film pressures, all films examined were considerably more expanded than similar films at the air-liquid interface. At high pressures, however, the film characteristics approximated those of the air-liquid interface.

[18] Askew & Danielli. Proc. Roy. Soc. London A 155: 695. 1936.
[19] Alexander, A. E., & T. Teorell. Trans. Farad. Soc. 35: 726–737. 1939.

It appears that the film at the liquid-liquid interface, far from being fixed and unchanging, undergoes a series of transitions from an expanded state (especially at low concentrations of surface active materials) to a solid condensed film possessing comparatively high surface (or interfacial) viscosity, exhibiting possibly non-Newtonian or plastic flow characteristics. The suggestion given by Alexander and Schulman[20] that a "rigid" interfacial film is necessary for emulsion stability is in line with this argument.

These experiments, drawn from a larger body of data, offer evidence that the liquid-liquid interface presents features analogous to those of the air-liquid interface,[21, 22] and that the mathematical models adopted are sufficiently similar to require no great change in logical procedure. Such, however, is not the case with the solid-liquid interface, and a number of expedients have been introduced.

Harkins,[23] in a review of his work over many years in this field, summarizes the generalizations which he and his co-workers have developed, primarily from the thermodynamic approach, that is, in terms of the energy changes involved. He emphasizes that care must be taken, when analyzing spreading phenomena, not to confuse fresh solid surfaces with those which have been given an opportunity to reach equilibrium with all components in the system, which distinguishes between his initial ($S_{b/a}$) and his final ($S_{b'/a'}$) spreading coefficients. In another connection, he shows the importance of the time factor in the generalization that, whereas some liquids will spread initially over water as duplex layers, in no instance is such a system stable, but ultimately transforms into a group of liquid lenses connected by a monolayer of the liquid on the water.

THE SOLID-LIQUID INTERFACE

Inasmuch as the interfacial energy at the solid-liquid interface cannot be directly measured, a variety of indirect approaches has been adopted. The oldest and most frequently mentioned is that of the angle of contact between the liquid and the solid surface. Unfortunately, both the experimental and theoretical approaches to this topic have been clouded with difficulties, and, even today, there is no general agreement among investigators on all aspects of the problem.

---

20 Alexander, A. E., & J. H. Schulman. Trans. Farad. Soc. 36: 960–964. 1940.
21 Alexander, A. E. Ibid. 37: 117–121. 1941.
22 Schulman, J. H. Ibid. 37: 134–139. 1941.
23 Harkins, W. D. Intermolecular Forces and Two Dimensional Systems. Publication 21, Surface Chemistry. Am. Assoc. Adv. Science. 1943.

Even in the elements of the subject, difficulties of a semantic character exist.

Contact-Angle Relations. The geometry of the contact of a liquid with a solid may be systematized as shown in FIGURE 6. The first drawing represents the case where the liquid forms a finite contact angle with the solid. To the right of the lens, the substance will not spread in bulk but only as a monolayer—for example, a liquid-condensed film. The energy figures are arbitrary and are given to illustrate a possible case. The forces equivalent to the free energies of the system are drawn as vectors. In the second case, the contact angle is

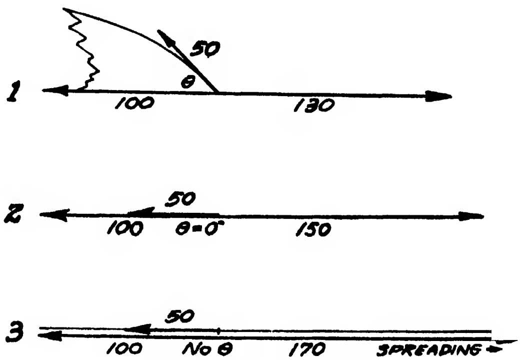


FIGURE 6 Three cases of liquid-solid "wetting"

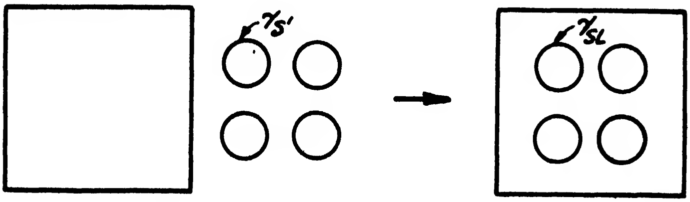
actually zero. Equilibrium is represented by numerically balanced forces. As may be anticipated, this instance of a contact angle which is precisely zero is rare. The third case is the one for which there is *no* line of contact, hence no contact angle, and the spreading continues to form a duplex film over the solid.[24]

The term, "zero contact angle," is employed very loosely in the literature. A finite contact angle represents an equilibrium between three forces along a line of contact. When one of these forces predominates, as in the third case, equilibrium cannot exist along any line of contact, and spreading of the liquid occurs in the direction of that force. This spreading does not represent a "zero contact angle," but rather no contact angle at all. Some assertions in the literature are based on the supposition that this spreading corresponds to a zero contact angle with an implied equilibrium of forces which, in truth, does not exist. Yet, for example, this assumed equilibrium is used to demonstrate the valid-

[24] Gans, D. M. J. Phys Chem. 49: 165–166. 1945.

ity of such controversial relationships as Antonoff's Rule, which holds only in the rare instances represented by the second case above.

The immersion of a solid in a liquid may be idealized as shown in FIGURE 7. In this system, f_I is the free energy change per cm² on immersion of the spheres in the liquid represented by the squared area. When the new solid-liquid interface is formed, the free energy change is represented by the difference of $\gamma_{S'}$ (the free energy of the solid in equilibrium with the liquid vapor) and γ_{SL} (the free energy of the solid-liquid interface). The forces acting at the interface resolve to γ_L (the



$$-f_I = \gamma_{S'} - \gamma_{SL}$$
$$= \gamma_L \cos \theta$$
$$-f_I \begin{cases} + \text{ for } \theta < 90° \\ - \text{ for } \theta > 90° \end{cases}$$

FIGURE 7. Diagram representing immersion of solid into liquid.

surface tension of the unchanged liquid) multiplied by the cosine of θ (the angle of contact). The term, "wettability," has been used in the literature to describe some quality associated with the spreading of a liquid over a solid. This term may be defined in terms of the free energy change of immersion. Qualitatively, on this basis, a contact angle greater than 90° indicates that the solid prefers to remain "unwetted," while, at angles less than 90°, the solid tends to be "wetted" by the liquid.

This view is implicit, but not expressed, in many treatments of the classical problem of contact-angle measurements. In a recent paper by Irons,[25] experimental data are given for contact angles of a variety of liquids on different metals as determined by a pressure differential method. The method apparently circumvented the difficulties of contamination noted for other methods, such as the tilting-plate method.

[25] Irons, E. J. Phil. Mag. 34: 614–625. 1943.

Irons found that, within experimental error, the liquids studied spread over the walls of the capillaries selected.

The difficulties in the measurement of contact angles are emphasized in the work of Cassie and Baxter[26] on the wetting of porous and irregular surfaces. The grid structure of some fabrics and animal structures, such as the feather, give, in effect, a larger *apparent* advancing contact angle than would be expected from a plane surface of the same material.

While the literature on contact-angle measurements is abundant, and some of the work is painstaking, there are strikingly few data which can be accepted without reservation. This suggests that the validity of the concept may indeed be questioned, and an alternative view, supported by considerable experimental evidence, can be offered on the simple basis of whether or not a liquid spreads over a solid surface. If this is accepted, the observed angle at which a liquid appears to meet a solid surface ceases to have any thermodynamic meaning and could be designated, instead, as the "angle of approach."

EXAMPLES FROM PRACTICE

It will be instructive to examine several important industrial processes from the standpoint of measurable interfacial relationships, noting the correlation which has so far been found and to supplement the experimental information with hypothesis.

Surface Wetting and Spreading. Wetting of a solid surface by a solution presents, in some respects, an attractive subject, for it is one of the most familiar and challenging of all. As an example, the recent improvement in water-repellence of fabrics has been of service to nearly everyone. For our discussion, however, it is the opposite effect, that of improved spreading which is of equal interest. A large industry—that of insecticide sprays—depends specifically on the ability of a suspension to spread uniformly over leaves and branches which may exhibit repellency.

In an interesting series of studies by Martin[27] an effort was made to correlate various physical properties with spray retention. It was found that the area of spread of droplets could be related to measurable contact angles and spreading coefficients. For these experiments, a variety of surfaces was employed, including paraffin, cellulose nitrate and acetate, and shellac. The last was the least reproducible. On the basis of several hundred observations, Martin found that the area of

---

[26] **Cassie, A. B. D., & S. Baxter.** Trans. Faraday Soc. **40:** 546–551. 1944.
[27] **Martin, H.** J. Pomology **18:** 34–51. 1940.

spreading could be correlated with the advancing contact angle in 86.5%, and the equilibrium contact angle with 84.8% of the cases studied. It was found, further, that the relations observed held for surface-active compounds of diverse chemical structure.

Penetration. Penetration of a capillary or network of a solution is another process presenting a complex of properties. Two cases are immediately evident: (a) those in which the liquids spread over the capillary surface, and (b) those in which the contact angle (or the "apparent contact angle") is sufficiently large to impede the advance of the liquid. Washburn[28] showed that the rate with which a liquid penetrates a capillary is directly proportional to the capillary radius, the surface tension of the liquid, the cosine of the angle of contact, and inversely proportional to the viscosity of the liquid.

$$l^2 = \left(\frac{\gamma \cos \theta}{\eta\, 2}\right)$$

$$\frac{dl}{dt} = \frac{r \cdot \gamma \cos \theta}{\eta\, 4\, l}$$

where l = length of liquid in capillary
t = time
r = capillary radius
η = viscosity
γ = surface tension
θ = angle of contact.

The interdependence of all factors is well shown in the Washburn formula. It is the expression of a rate process and is applicable, accordingly, to modern manufacturing methods based on continuous operations in which the time cycle is short and fixed by mechanical factors. Where a finite time is required for establishing a steady surface tension, or where selective adsorption occurs (particularly in dilute solution), the Washburn formula fails to predict performance adequately.

The quantity, $\frac{\gamma \cos \theta}{2\eta}$, was defined by Washburn as the "coefficient of penetrance or penetrativity" of the liquid, and is a measure of the velocity of penetration. Since the contact angle, θ, enters into the relationship, the material of the capillary determines, along with viscosity and surface tension, the rate of penetration. For liquids which do not wet the capillary wall ($\theta > 90°$), there is no penetration. Mercury in a glass capillary is an example of this case. For liquids which spread

[28] Washburn, E. W. Phys. Rev. 17: 273–283. 1921.

over the solid surface, or where the contact angle would be designated as "zero," the penetrativity is reduced to the ratio of surface tension to viscosity.

It will be inferred that penetration into a capillary mass, such as that of wood, under steeping or soaking conditions, is evidently aided by a high rather than a low surface tension. Addition of surface-tension depressants, consequently, is of no practical value as demonstrated in the experiments of Stamm and Petering,[29a] and McGovern and Chidester.[29b]

A loose meshwork or reticulum of fibers, on the other hand, presents a different condition. Here, the fibers are spaced so widely that the contribution to wetting made by capillary rise is at a minimum (although not necessarily zero) and the condition is that of very small solid areas for which wetting by spreading is the major factor. If, in such a skein, felt, or batting, the contact angle is appreciable, wetting of the fibers is hindered or prevented completely, and a surface-tension depressant then facilitates penetration of the mass.

This appears to be the case of the widely used Draves[30] and Canvas-disc[31] tests. Initial immersion of the skein of raw cotton results in the solution wetting and penetrating into the separate threads comprising the skein. The air trapped between the threads is gradually displaced, along with some of the natural oils present, and the buoyancy of the skein decreases to the point where the weight pulls it down. In the canvas-disc test, the liquid penetrates the fabric displacing air and oil until the disc sinks under its own weight.

Both of these tests evidently present four interfaces: air-solid, air-liquid, liquid-liquid, and solid-liquid. Since the viscosity (that of the liquid at the air and at the solid interfaces as well as in bulk) is presumed to be constant, the surface tension and the angle of contact of the liquid on the solid appear to be the main factors (FIGURES 8 and 9). It is, perhaps, surprising to find as much regularity as is shown in the measurements of the contact angles of these solutions against clean paraffin blocks. These measurements were made under carefully standardized experimental conditions and, while no claim for inherent accuracy is made, they undoubtedly represent a series of relative values for comparable, non-equilibrium conditions. A rough correlation with the wetting properties of the solutions, analogous to that noted by Martin,[27] can be made.

[29a] **Stamm, A. J., & W. K. Petering.** Ind. Eng. Chem. **32:** 809–813. 1940.
[29b] **McGovern, J. N., & G. H. Chidester.** Paper Trade Jour. 111 (24): 35–38. 1940.
[30] **Draves, C. Z.** American Dyestuff Reporter (Proc. AATCC) **28:** 421–428. Aug. 7, 1939.
[31] **Seyferth, H., & O. M. Morgan.** Ibid. **27:** 525–532. Sept. 19, 1938.

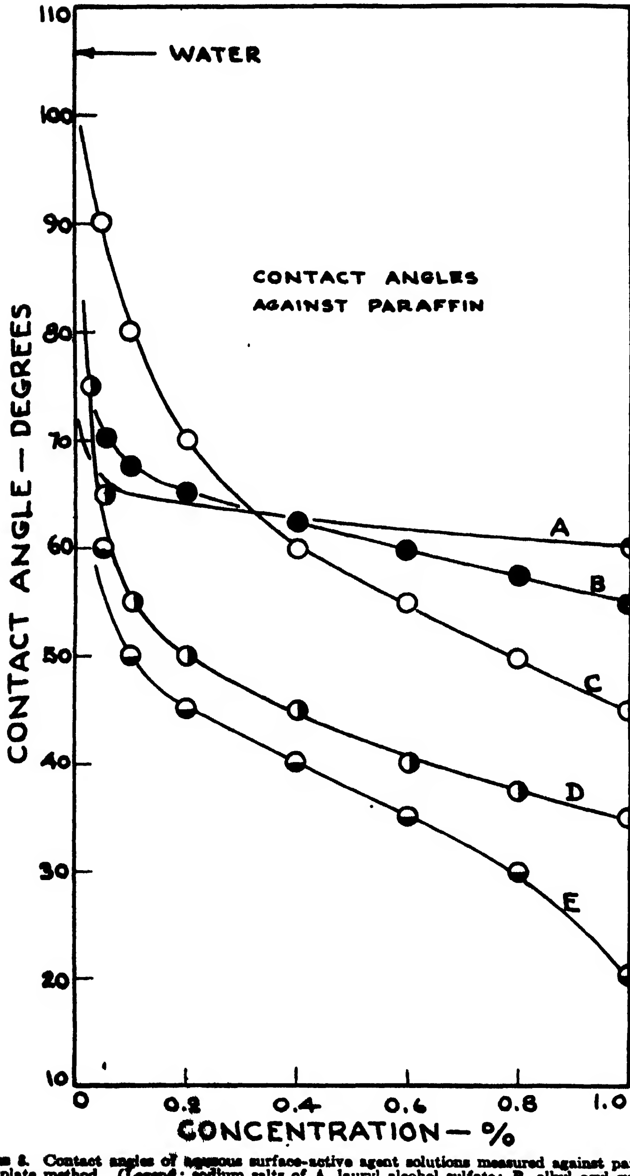


FIGURE 3. Contact angles of aqueous surface-active agent solutions measured against paraffin by tilting-plate method. (Legend: sodium salts of A. lauryl alcohol sulfate; B. alkyl-aryl sulfonate; C. Secondary alcohol sulfate; D. alkyl-aryl sulfonate; E. octyl alcohol ester sulfosuccinic acid.)

It is not always appreciated that fabric-sinking tests can be made over a relatively restricted range. Thus, if the Draves sinking time

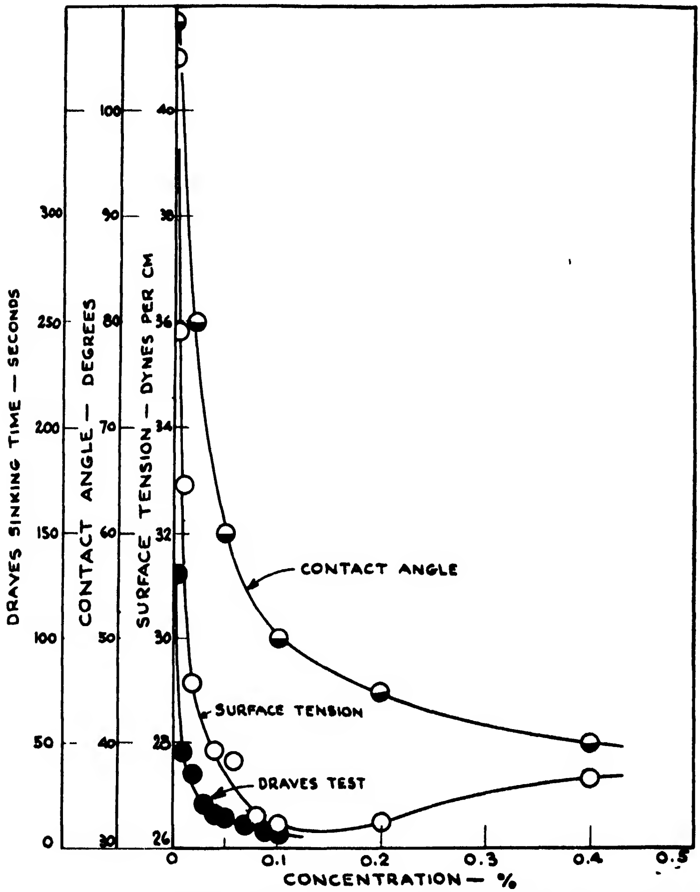


FIGURE 9 Curves showing Draves' test sinking times, surface-tensions values, contact-angle measurements in relation to concentration of reagent (sodium salt octyl ester sulfosuccinic acid).

exceeds about five minutes or is less than about six seconds, reproducibility of the measurement becomes insufficient to provide significant data. All practical sinking times appear to fall into the concentration region where the surface-tension-concentration curve has leveled off, and it is not likely that any of the apparent anomalies men-

tioned previously have significant bearing on fresh solutions of surface-active material. Where a solution is used for processing a web or a series of units of a porous material, however, the surface-active substance is preferentially adsorbed at the solid-liquid interface result-

TABLE 1. CHANGE IN PROPERTIES OF SOLUTIONS OF SURFACE-ACTIVE AGENTS ON EXHAUSTION

(Draves' Test procedure; 5-gram raw cotton skeins; 6-gram sinker; Temperature 30° C.)

Number of Immersed Skein	Sinking Time (Seconds)	Surface Tension (Dynes/cm.)	Contact Angle (paraffin) ° Arc
Sodium salt of octyl ester sulfo-succinic acid; concentration 0.015%			
0		29.3	53
1st	20.2		
5th		30 8	50
6th	29.6		
10th		31.5	55
11th	53.5		
15th		31 8	57
17th	101		
20th		32.4	60
21st	120		
25th		33.4	63–4
26th	210		
30th	∞	35 8	68
Sodium salt of lauryl alcohol sulfate; concentration 0.1%			
0		30.6	65
1st	12		
5th		29.6	65
6th	13.4		
10th		28.8	65
11th	13 6		
15th		29 0	65
16th	16.3		
25th		29.2	55
26th	15.6		
35th		29.9	50
36th	25.8		
45th		31.9	55
46th	60	32.4	65

ing in exhaustion or a differential extraction of the solution components. This result is frequently observed in dye baths. When such extraction occurs, the gradual depletion of surface-active substances leads to prolonged wetting times.

An experimental demonstration of exhaustion and the changes in surface tension and contact angle are given in the data of TABLE 1 and FIGURES 9 and 10. In this series, 5-gram skeins of raw cotton were

immersed in the wetting bath and then wrung out in a wringer with the expressed solution falling back into the bath. Draves sinking times were measured and surface tension values and contact-angles of the solution against paraffin were determined as the exhaustion continued. It will be observed that the surface-tension readings in TABLE

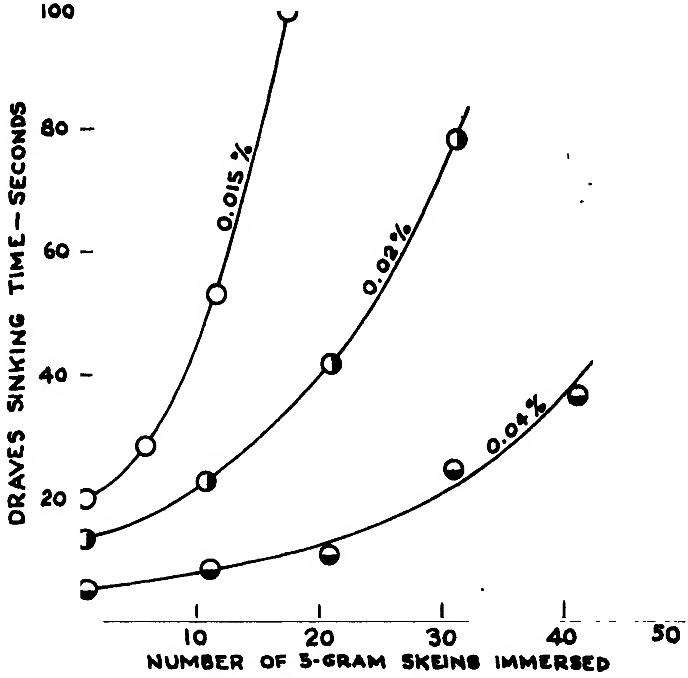


FIGURE 10. Increase in Draves' wetting times as a function of number of cotton skeins immersed per liter of solution. (Sodium salt of octyl ester sulfosuccinic acid; temperature 30° ± 1° C.; 6-gram hook used for Draves' test.)

1 show an unmistakable trend toward higher values as the surface-active material was extracted from solution. Contact-angle measurements against paraffin for the same solutions did not show a comparable regularity, for the imperfections of this measurement preclude data of the necessary precision.

The one experiment described was performed at a concentration of 0.015% with the sodium salt of the sulfonated octyl ester of succinic acid. Higher concentrations of this and other reagents (fatty alcohol sulfates, alkyl-aryl sulfonates) showed that the surface tension changed

slowly at first and then rapidly as the extraction progressed. The larger quantity of reagent present in these solutions allowed as many as 40 skeins to be processed per liter of solution before the sinking time reached one minute. In these experiments, a minimum in the surface-tension curve was noted, but the sinking time of around 15 seconds remained approximately constant. Since the immersion and wetting of the skein displaces oil present on the cotton fiber, and the solution then emulsifies the oil, it is reasonable to consider the process as follows: the surface-active material is present initially in sufficient quantity so that a reservoir, possibly in the form of micelles, exists; on continued immersion of the fibers, displaced oil emulsifies and "dissolves" part of the surface-active species; at a point where the surface-active material is only slowly released, or where the total available is insufficient to allow adequate spreading and wetting of the solid surface, sinking times lengthen, and the surface tension rises markedly. Exhaustion is most rapid, of course, where the surface-active species is no longer available from a micellar reservoir. Foaming is also greatly diminished at the point where exhaustion becomes most rapid.

This hypothesis suggests that surface and interfacial tension values, as well as contact-angle data, offer an intensity measure, while the sinking times for skeins or fabrics, provide a measure of capacity of the wetting qualities of a processing bath.

An analogous case is the process for impregnation of cotton batting with rubber latex.[32]

When the cotton batting is immersed in the latex bath, the aqueous solution forming the continuous phase wets the fiber surface, and spreads rapidly. The latex particles follow the stream and encounter the matted fibers at the batting surface which act as a coarse filter. Mechanical working of the batting between screens serves to aspirate the latex into regions in the interior of the fiber matrix or where the fibers are more closely matted. The adhesion of the serum to the hydrophilic fiber is high. The adhesion of the latex particles to the fiber covered with the aqueous film is nil. When the saturated batting, entraining several times its weight of latex, is subjected to compression in squeeze rolls, surplus liquid is expressed, and the latex particles, being mobile and non-flocculated, are free to follow the flow of the expressed liquid, finding no greater hindrance to leaving the web than in entering. The serum, however, remains tenaciously adherent to the surfaces.

The result is a differential extraction, with the non-rubber compo-

[32] Fischer, E. K. Textile Research. 14: 333–341. 1944.

nents, such as proteins, resins, and added surface-active agents, extracted at a rate greater than the rubber. A comparative enrichment

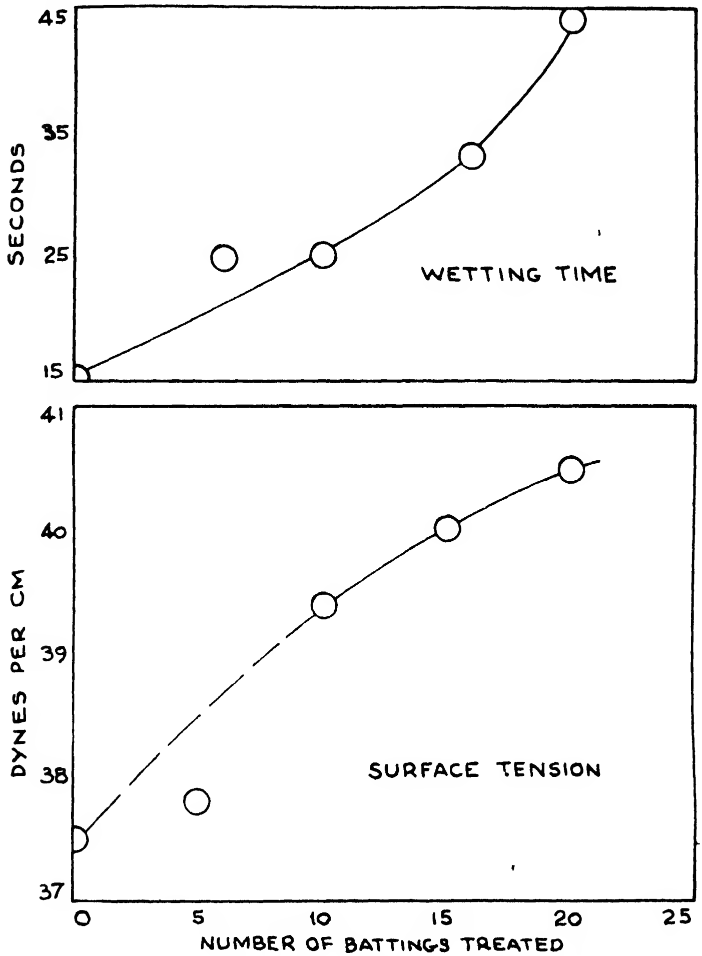


FIGURE 11. Change in wetting times and surface-tension values on bath depletion.

of rubber in the bath and depletion of the water-soluble materials ensues, accompanied by a gradual increase in wetting time and surface tension (FIGURES 11 and 12).

This effect is illustrated in FIGURE 13, where the wetting qualities of the bath are shown as a function of the squeeze-roll pressure and the solids retained by the batting. The quantity, "% of theoretical solids," was obtained by the following computation:

$$\%\ \text{Theoretical Solids} = \frac{100W_D}{W_L}$$

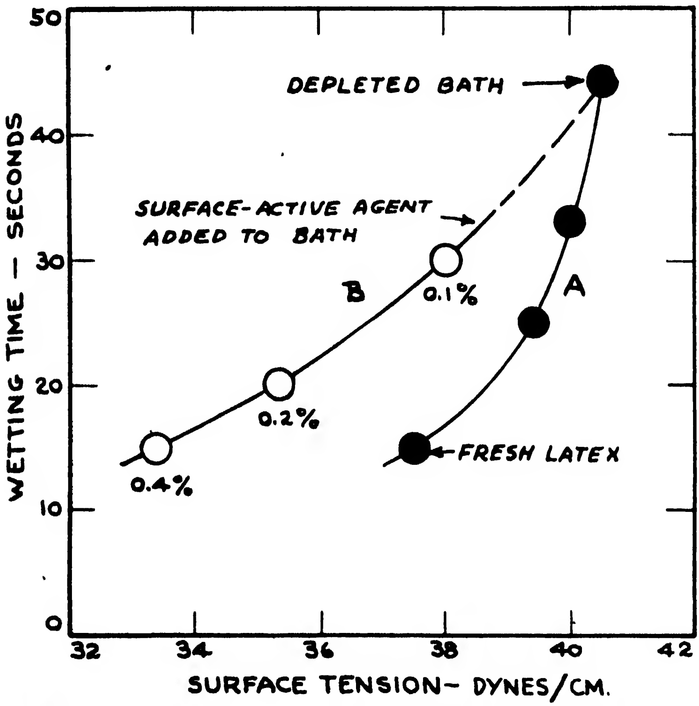


FIGURE 12. Depletion of latex bath and partial restoration of wetting qualities by addition of a surface-active agent. (Napper cotton batting; 80% latex; sodium salt of mixed fatty alcohol sulfates as wetting agent.)

where W_D is the added solids in the dried sheet and W_L is the theoretical solid content computed from the non-volatile content of the latex and the total latex absorbed in the wet batting. It will be noted that the latex baths of the higher surface-tension values retained a larger proportionate amount of the latex solids and, at low squeeze pressures, approached values from 90 to 96%. At lower surface-tension levels, the maximum retained was around 80%. As the squeeze pressure was increased, there was a striking trend in all cases to lower percentages.

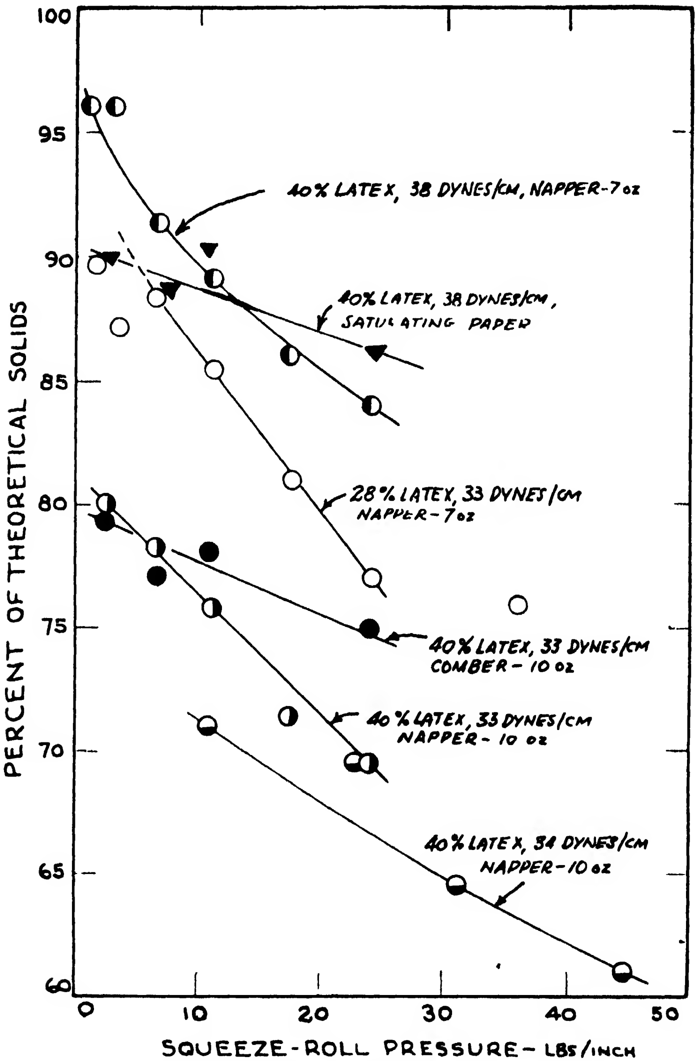


FIGURE 18 Differential extraction of serum and rubber solids for various latices and base materials as a function of squeeze-roll pressure, latex concentration, and surface tension of bath

On the other hand, if the serum does not effectively spread over the fiber surface, mechanical working of the batting aspirates the latex

and minute pockets or lakes of the liquid are formed throughout the fibrous reticulum. Expression of the excess saturant takes place with the serum withdrawing from the fibers as water does on an oiled surface, carrying with it the full complement of latex particles.

In either case, however, drying the impregnated matrix of fiber and latex results in liquid streaming within the matrix spaces toward the surface where evaporation occurs. Some of the latex particles are carried with the streaming serum, and the concentration of rubber is increased above that in the central interstices of the batting. High volume of the liquid phase (low solids content) obviously increases the quantity of latex particles transported, and the surface layering then becomes so great that the sheet is left completely devoid of rubber solids in the center. A sheet of this structure splits readily, because only the matted fibers, uncemented by adherent rubber, provide mechanical bonding With a minimum of water retained by the web, represented by a latex impregnant of high solid content, there results a minimum of migration for the additional reason that the latex coagulates when a relatively small volume of the water evaporates. and the coagulum is then trapped in the interstices of the web, unable to migrate, while the residual water evaporates.

An ingenious explanation of the effect of surface active materials on the biological activity of phenols, which takes into account penetration rate, surface-tension effects, and the region of micelle formation, is given by Trim and Alexander.[33]

At a fixed concentration of hexyl resorcinol, the rate of penetration into *Ascaris lumbricoides* increases with surface-active agent concentration to a maximum which corresponds with the lowest interfacial tension. With further increase in soap concentration, biological activity falls to zero, and the interfacial tension value of the hexyl-resorcinol-soap mixture rises to that of the soap alone, although, at lower soap concentrations, it showed a minimum. The shift in the interfacial tension values is taken to indicate the region in which micelle formation begins. The hexyl resorcinol is distributed between the interface and the micelle. At high soap concentrations, the hexyl resorcinol is largely locked up in the micelles and the mixture shows a negligible biological activity.

Foam Stability. Recent work on foam stability has not fully taken into account the role of surface viscosity as a factor second only to surface-tension relationships. Plateau,[34] in 1870, noted that the forces

---

[33] Trim, A. R., & A. E. Alexander. Nature 154: 177–178. 1944.
[34] Plateau, J. J. Pogg. Ann. 141: 45. 1870.

which controlled foaming were surface viscosity in relation to surface tension. Plateau used a magnetized needle floating on the surface to show that the surface viscosity of aqueous saponin solutions is different in magnitude from the viscosity of the liquid in bulk.

Surface-viscosity measurements on fatty acids and alcohols, using highly refined techniques, reported by Harkins and his collaborators,[35, 36, 37] showed that liquid films exhibited Newtonian flow characteristics, with the logarithm of surface viscosity proportional to film pressure, and that the surface viscosity of such liquid films increases with length of the hydrocarbon chain. Plastic viscosity, in which the measured viscosity varies with rate of shear, was noted for high pressure-condensed films, and phase transitions in the films could be detected by surface viscosity measurements. In general, the surface viscosities of monolayers of acids were considerably less than those of the corresponding fatty alcohols.

Foam stability of commercially practical systems has been measused by many investigators using methods based on ability to form a foam, the rate of foam drainage, the life and volume of the "dry" film, etc.[38, 39] It has been found that increasing the concentration of the surface-active material from 0.05% to 0.10% increased the foam stability by a factor in excess of tenfold. With sodium oleate at a concentration of 0.04%, foam stability was nil. Addition of pectic substances, gums, and organic compounds increased stability. Of great importance is the pH region at which soaps, in particular, are maintained The presence of free fatty acid has a role in foam formation, and available data show that unstable foams may be stabilized by the addition of free fatty acid as well as by fatty alcohols.

It is unfortunate that data on surface viscosity and surface tension have not been obtained along with measurements on foam stability. However, it is possible to suggest an hypothesis qualitatively connecting the factors from evidence at hand: In the process of foam formation by agitation with air or an inert gas, the formation of a large area of liquid-air interface is facilitated, both by low surface tension and low bulk viscosity, especially since the mechanical work performed in the process is comparatively small. Once the foam is formed, drainage is slowed by a high bulk-liquid viscosity, and, as the drainage progresses, the effect of high surface viscosity assumes a paramount role

[35] **Harkins, W. D.** J. Phys. Chem. **42:** 897–910. 1938.
[36] **Boyd, E., & W. D. Harkins.** J. Am. Chem Soc. **61:** 1188–1195. 1939.
[37] **[illegible]tting, G. C., & W. D. Harkins.** Ibid. **62:** 3155–3161. 1940.
[38] **[illegible]oss, J., & G. D. Miles.** Oil and Soap **18:** 99–102. 1941; J. Phys. Chem. **48:** 280–290. 1944.
[39] **Merrill, R. C., Jr., & F. T. Moffett.** Oil and Soap **21:** 170–175. 1944.

in preventing collapse of the film. Adsorbed surface-active substances are further concentrated in the film, forming a condensed oriented film which has comparatively high rigidity, exhibiting, undoubtedly, plastic-flow properties. Mechanical agitation (e.g. vibration), in some cases, may supply sufficient force to exceed the yield value of the film, resulting, then, in gradual collapse.

Non-Aqueous Systems. While surface active agents have their greatest and most spectacular utility in aqueous systems, a large num-

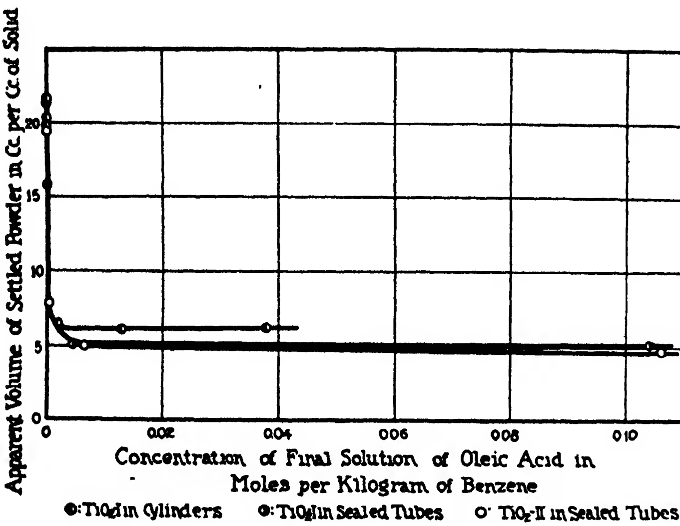


FIGURE 14. Variation of extent of settling of titanium dioxide with amount of oleic acid present.

ber of industrial products based on non-aqueous media are modified by their use. At present, it is impossible to draw on a fund of experimental evidence for elucidation of all the observed phenomena, but enough data are available, particularly with solid-liquid systems, to offer a guide. The experimental techniques which have been employed to advantage are sedimentation equilibria and rheological measurements.

If a powdered solid such as a pigment is shaken with an organic liquid like benzene and then allowed to settle, the rate of settling and final sedimentation value indicate the extent to which the dispersed solid is flocculated. Generally, solids which settle to low equilibrium volumes do so very slowly. Those which settle rapidly exhibit relative-

ly high final volumes. A study of these factors[40, 41, 42] showed that many of the observed relations could be explained by the adsorption from solution of surface-active substances in a unimolecular film.

The presence of such a unimolecular film on the solid resulted in small settling volumes, and the method was found sufficiently sensitive to compute the interfacial area of the solid using a value for the adsorbed substance (fatty acid) obtained by other means. In these experiments, water was found to prevent deflocculation with oleic acid on

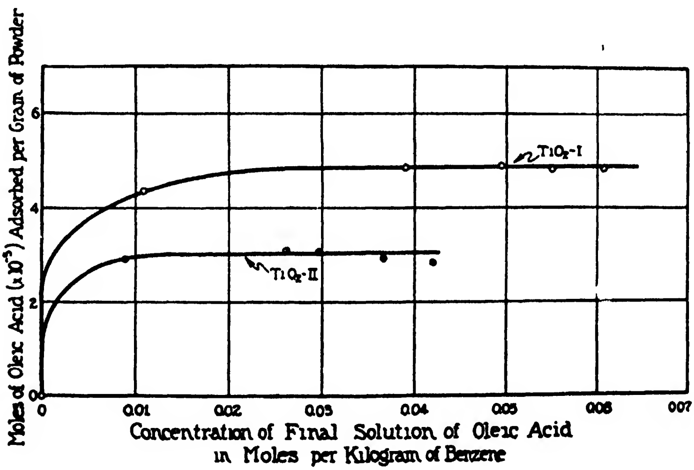


FIGURE 15. Adsorption curves for dried titanic oxides

titanium dioxide particles, a result attributed to the preferential adsorption of water at the hydrophilic pigment surface (FIGURES 14 and 15).

In solid-liquid suspensions, in which the solid content by volume is high (e.g., 15–30%), settling is virtually impossible under gravity, but the effects of flocculation are evident in the rheological characteristics of the dispersion. Plastic flow with a measurable yield value and, in many cases, marked thixotrophy are evident. The changes in yield value of such systems (measured on a rotational viscometer) have been taken as the criterion of the influence of surface-active agents.[43]

[40] **Ryan, L. W., W. D. Harkins & D. M. Gans.** Ind. Eng. Chem. **24:** 1288. 1932.
[41] **Harkins, W. D., & D. M. Gans.** J. Phys. Chem. **36:** 86–97. 1932.
[42] **Harkins, W. D., & D. M. Gans.** J. Am. Chem. Soc. **53:** 2804. 1931.
[43] **Fischer, E. K., & C. W. Jerome.** Ind. Eng. Chem. **35:** 126–130. 1943.

Data illustrating the lowering of yield value are given in TABLE 2. The microscopic appearance of flocculated pigment dispersion in comparison with the same composition substantially deflocculated is shown in PLATE 1.

TABLE 2. CHANGE IN YIELD VALUE OF PIGMENT DISPERSIONS ON ADDITION OF SURFACE-ACTIVE AGENTS

Reagent	Yield Value (Dynes/cm²)
Carbon black (9.7 volume % in glycerol)	
None (control)	3800
Sodium salt sulfonated octyl ester succinic acid	2000
Sodium salt lauryl alcohol sulfate	1200
Sodium salt secondary alcohol sulfate	820
Barium lithol toner (19.8 volume % in glycerol)	
None (control)	1120
Sodium salt lauryl alcohol sulfate	220
Sodium salt alkyl-aryl sulfonate	0
Ultramarine blue (32 volume % in mineral oil)	
None (control)	1900
Sodium salt secondary alcohol sulfate	1400
Sodium salt sulfonated octyl ester succinic acid	530
Lecithin	0

The influence of trace quantities of water and the extent to which surface-active materials prevent a large increase in yield value are shown in TABLE 3.

The manner in which deflocculation is accomplished in such solid-liquid suspensions or dispersions is a subject still incompletely explored. The theory is founded on the following experimental findings: (a) change in rheological properties; (b) insignificant changes in measurable surface-tension effects of the liquid at the air interface on the addition of surface active substances; (c) demonstrable changes in the energy relation at the interfaces through studies on heats of immersion;[44, 45] (d) the relatively greater quantities (in comparison with aqueous systems) of surface-active materials necessary to obtain marked alteration in

[44] Harkins, W. D., & R. Dahlstrom. Ind. Eng. Chem. 22: 897–902. 1930.
[45] Boyd, G. E., & W. D. Harkins. J. Am. Chem. Soc. 64: 1190–1204. 1942.

TABLE 3. CHANGE IN YIELD VALUE OF "HYDROPHILIC" PIGMENT DISPERSIONS IN MINERAL OIL ON ADDITION OF WATER (3% ON PIGMENT)

Pigment	Reagent	Initial Yield Value (Dynes per cm²)	Yield Value after Water Addition (Dynes per cm²)
Titanium dioxide[a]	None (control)	*c*	*c*
Titanium dioxide	Lecithin	1300	1500
Titanium dioxide	Sodium salt octyl ester sulfosuccinic acid	3600	3200
Titanium dioxide	Zinc naphthenate	2900	4400
Ultramarine[b]	None (control)	1030	*c*
Ultramarine	Lecithin	0	92
Ultramarine	Zinc salt octyl ester sulfosuccinic acid	300	850
Ultramarine	Octyl amine salt of alkyl-aryl sulfonic acid (?)	170	9200

a 25.8% by volume
b 28.2% by volume
c Too high to measure

properties, suggesting a role comparable to that of protective colloids in aqueous systems; (e) the possible existence of micelle structure;[46] and (f) the influence of adsorbed moisture (and other substances) on the solid before incorporation into a non-aqueous medium.[47]

Such questions as the relative stability of a dispersion involve the magnitude of the free energy of the surfaces as well as the contact angle between them. The trend is summarized in the generalization that the dispersion of a powder in a liquid is more stable than unmixed liquid and powder for contact angles under 90° but that the system will resist the mixing operation for contact angles over 90°.[48] As another illustration, a paint will spontaneously resist the painting operation for all angles of contact above 0°.

In this review, an effort has been made to connect knowledge of the theoretical aspect of interfacial activity with the use of surface-active agents. While many applications remain highly empirical—for lack of sufficient information—it is possible, nevertheless, to employ the results of scientific studies in the explanation of the effects of surface-

[46] Lawrence, A. S. C. Trans. Farad. Soc. 34: 660. 1938.
[47] ———, & H. G. Van Selms. Rec. Trav. Chim. 62: 398–426. 1943.
[48] ———, & D. M. Gans. Dispersion of Finely Divided Solids. Chapter in Colloid Chemistry 6. Edited by Jerome Alexander. Rheinhold Publishing Co. New York, N. Y. In press.

active agents. It is necessary, in particular, to recognize that no single physical measurement will suffice for complete understanding or for laboratory control of an industrial process. This subject constitutes, in fact, an attractive field for research in the borderline between fundamental studies and practical applications.

DISCUSSION OF THE PAPER

Dr. E. I. Valko (*Onyx Oil and Chemical Co., Jersey City, N. J.*):

The problem of the correlation between interfacial tension and stability of emulsions was mentioned in the paper by Dr. Fischer and Dr. Gans. There is a fundamental difficulty in the establishment of such a correlation: the stability of emulsions is not a thermodynamic problem but one of kinetics. If the stability were a thermodynamic problem, the interfacial tension of the emulsified material against the continuous phase would alone determine the stability. However, emulsions are, in general, thermodynamically unstable systems and what we call stability is, in fact, the slowness with which the system approaches the state of equilibrium, namely, the complete separation. The rate of creaming of the emulsion is not proportional to the energy gained by reducing the interface which is measured by the interfacial tension, but is determined by the activation energy of the coalescence of the oil globules. This makes it understandable that, for instance, the electric potential of the particles is an important factor for the "stability" of emulsions independently of the interfacial tension, notwithstanding the fact that the electrical energy is a part of the interfacial tension; and it explains also the fact that the viscosity or the mechanical strength of the surface active film is another important factor. While, theoretically, it is possible that the interfacial tension has a minimum at a certain value of the size of the particles, and, thus, an emulsion is stable in a true thermodynamic sense, no emulsion has been proved yet to be of such a nature and most of the emulsions used in industry definitely do not belong in this class. The only known exceptions are micellar systems containing not substantially more than one part of weight of insoluble material to one part of surface active emulsifying agent. These systems, in which the water insoluble material is dissolved in the micelles, are optically clear and represent a thermodynamically stable, reversible equilibrium. They are generally designated as solutions and not as emulsions.

Dr. Kurt Wohl (*Department of Chemical Engineering, University of Delaware, Newark, Del.*):

For several solutions of long chain electrolytes (soaps) anomalies of electric conductance, surface tension and osmotic pressure have been found which are illustrated in FIGURE 1: The osmotic coefficient is defined as $g = P/cRT$ (P = osmotic pressure). The ordinate scales are arbitrary; the abscissa scale indicates roughly the range in which these anomalies occur; and the mutual position of the three curves with respect to the abscissa corresponds roughly to experiments. According to Gibbs' equation, the part of the σ curve with a positive slope seems to require a negative surface adsorption while McBain proved that adsorption is positive. This situation is known as McBain's paradox (McBain and Mills, 1939[1]).

An explanation of the anomaly of conductance has been given by Hartley and Murray (1935).[2] It was based on the assumption that in the solution a micella of the type $A_n K_n$ is formed (A = long chain anion, K = small kation), which is in equilibrium with free A- and K- ions, while the concentration of intermediate products is negligible. They also brought the surface tension anomaly in connection with micella formation, without, however, treating the latter point in a conclusive way. The same may be said of Powney and Addison (1937),[3] Alexander

[1] McBain, J. Wm., & G. F. Mills. Rep. on Progress in Physics 5: 20. 1939.
[2] Hartley, G. S., & B. C. Murray. Trans. Far. Soc. 31: 183. 1935.
[3] Powney, J., & C. C. Addison. Ibid. 33: 1252. 1937.

(1941–42)[4] and Cassel (1942).[5] Other explanations were given by McBain and Mills (1939),[1] Long and Nutting (1941)[6] and Hauser (1942).[7]

We want to show quantitatively that the minimum and the maximum of the surface tension curve are a natural consequence of micella formation. Two points are essential. First: If the law of mass action is applied to the equilibrium between the micella $A_{20}\,K_{10}$ and the single ions, it appears that the concentration [A] of the free anion A has a maximum as is shown in FIGURE 2, in which a certain "reduced" unit has been used for concentrations (the product $[A] \times [K]$, of course, steadily increases with increasing concentration). Second: It is well known from the theory of Nernst's electrochemical potential and the Zeta potential that ions of one charge can be adsorbed at a surface, while the ions of opposite charge are not adsorbed. Thus, for equilibria with surfaces, the neutrality condition is not as strict as it is for equilibria between two bulk phases. (If it were not for the neutrality condition the two products of dissociation of a molecule AK could, of course, be separated by selective solubility in another bulk phase.)

The selective adsorption is treated thermodynamically by introducing for the anion a high coefficient of adsorption, k_A (k_A = limit of $\frac{n_A}{[A]}$ for small n_A; n_A Fk_{AK} $\left(k_{AK} = \text{limit of } \frac{n_{AK}}{[A][K]}\right)$ for small n_{AK} = moles of adsorbed anions per cm.²) and, for the monomolecular salt AK, a low coefficient of adsorption. We choose such values of k_A and k_{AK} that, up to soap concentrations somewhat beyond the maximum of [A] (FIGURE 2), the ion A is practically the only particle which is adsorbed. In this range of soap concentration, surface tension then depends only on the concentration of A. As the latter moves backwards to smaller concentration, when the soap concentration is increased beyond the maximum of [A], surface tension must move backwards with it according to Gibbs' equation, i.e., it must pass through a minimum at the maximum point of [A] and then increase. If the soap concentration is increased further, salt adsorption starts, so that the surface tension curve becomes normal again after having passed through a maximum. The σ-curve of FIGURE 1 is thus qualitatively explained.

The mutual position of the anomalies of the three curves of FIGURE 1 can be understood with the help of FIGURE 2. The minimum of σ (FIGURE 1) is fixed by the maximum of [A] (FIGURE 2). The anomalous drop of the osmotic coefficient, as it follows from our assumptions, is shown in FIGURE 2, too. It can be seen that the concentration, at which 50% of the drop has occurred, is higher than the concentration at the maximum of [A]. Therefore, the drop of the osmotic coefficient in FIGURE 1 must occur at a higher concentration than the minimum of σ, as is the case. The steepest descent of the λ-curve is given by the point of greatest curvature of the $[A_{20}\,K_{10}]$ curve (Hartley and Murray). As the latter point occurs in FIGURE 2 at a concentration higher than the two other characteristic concentrations mentioned, the steepest descent of λ-curve must, in FIGURE 1, occur at the highest concentration, as is the case.

In order to give this explanation a quantitative form, van der Waals' equation was used for the adsorbed phase. We assumed the proper molecular area of anion A and of salt AK to be equal, and first neglected the mutual electric repulsion between the anions in the surface. With the appropriate choice of constants, we thus obtained curve I of FIGURE 3. The branch I — I′ gives the number of adsorbed anions, relative to the maximum number of adsorbed particles, as a function of the log of soap concentration in "reduced" units. The curve I — I gives the adsorption of anion + salt. The difference between the branches I and I′ shows, of course, the adsorption of salt alone. The important feature is the minimum in the curve I.

From the adsorption curve I the decrease of surface tension ($\sigma_0 - \sigma$) follows according to van der Waals' equation of state for the adsorbed molecules or ac-

[4] Alexander, A. E. Nature 148: 752. 1941; Trans. Far. Soc. 38: 54. 1942.
[5] Cassel, H. M. J. Am. Chem. Phys. 10: 246. 1942.
[6] Long, F. A., & G. C. Nutting. J. Am. Chem. Soc. 63: 625. 1941.
[7] [illegible]. A. Advances in Colloid Science 1: 391. 1942.

cording to Gibbs' equation. Curve I of FIGURE 4 shows the result. Obviously, the general appearance of the empirical surface tension curve (FIGURE 1) is reproduced by this theoretical curve. The ordinate of FIGURE 4 contains, according to our procedure of calculation, a factor b which is the proper mole area of adsorbed soap. If we choose the proper area of one molecule of soap as 25Å^2, $\sigma_0 - \sigma$ for the minimum becomes, at room temperature, equal to 40 Dynes/cm which is the right order of magnitude. The branch I′ shows the surface tension which is due to the adsorption of the anion alone.

In order to check the effect of the electric charge, we introduced a potential energy of repulsion between adsorbed anions in contact with each other which was made somewhat larger than the potential energy of two elementary charges at a distance of 5 Å in water. With two different choices of the adsorption coefficients, k_A and k_{AK}, we thus obtained the curves II and III in FIGURES 3 and 4. It appears from FIGURE 3, that the charge tends to decrease the anomaly of adsorption, because the repulsion between the anions decreases anion adsorption and leaves more area for salt adsorption. In curve III, the minimum of adsorption has just disappeared. Still, because of the additional surface pressure due to the repulsion between electric charges the corresponding surface tension curve shows a minimum, though a weak one. The experimental curve seems to lie between curve I and II.

A closer adaptation of the theoretical curve to experimental data clearly requires several refinements of the assumptions made. A detailed account of the experimental data and the method of calculation used will be published elsewhere.

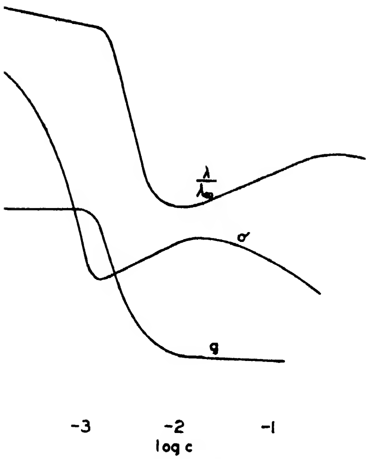


FIGURE 1.

λ = equivalent conductance
λ_∞ = equivalent conductance at infinite dilution
σ = surface tension
g = osmotic coefficient
c = total concentration of the salt AK in moles/liter

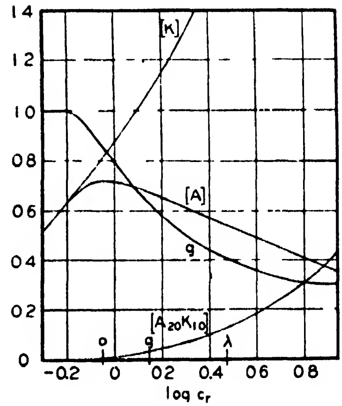


FIGURE 2.

[K] = concentration of the free cation
[A] = concentration of the free anion
$[A_{20}K_{10}]$ = concentration of the micella
c_r = total concentration of the salt AK
(all the above in "reduced units")
g = osmotic coefficient as it follows from our assumptions

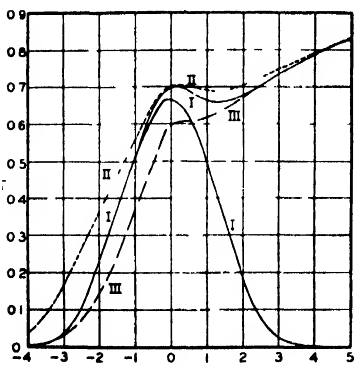


FIGURE 3.

n = number of particles adsorbed at the surface
n_{max} = maximum number of particles which have room at the surface
c_r as in Figure 2

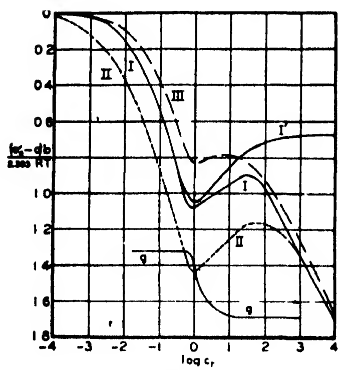
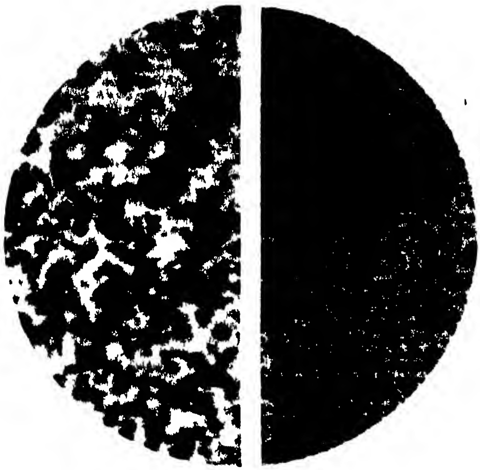


FIGURE 4.

σ_0 = surface tension of pure water
σ = surface tension of the solution
b = proper area of 1 mole of adsorbed particles
g and c_r as in Figure 2
The ordinate scale of g is not indicated.

PLATE 1

Deflocculated (right) and flocculated titanium dioxide dispersion in linseed varnish (photomicrographs x 800).



FISCHER AND GANS: SURFACE ACTIVE AGENTS AT INTERFACES

CERTAIN ASPECTS OF THE CHEMISTRY OF SURFACE ACTIVE AGENTS

BY DONALD PRICE*

Interchemical Corporation, New York, N. Y.

I. INTRODUCTION

Surface-active agents may be defined as substances which alter the energy relationships at interfaces. However, while such a definition may be quite correct, it is far too general to be of much practical value in the study of surface-active compounds.

We recognize surface-active agents as substances which, in dilute solution, are responsible for wetting, detergency, emulsification and related phenomena, and our interest lies in the relation between the aforesaid phenomena and the nature of the compounds which bring them about. We, therefore, find it more useful to speak of wetting-agents, detergents, emulsifiers, dispersing-agents and the like and to define each class separately in terms of performance rather than to attempt to define surface-active agents as a whole.

When we thus define surface-active agents on the basis of their performance, wetting-power, detergency, penetration, foaming-power, emulsifying-power and the like are regarded as properties of the said compounds. However, it should be pointed out that such "properties" are by no means simple, but represent exceedingly complex phenomena involving the operation of a number of factors simultaneously.

Moreover, we must not overlook the significant fact that the properties referred to above, which are displayed by surface-active agents, are closely interrelated, and a given surface-active agent usually possesses all of these properties to some degree. One function, however, generally predominates rather strongly over the others and thus forms the basis for the classification of the compound and for its selection for a particular use.

Compounds which display surface-activity are by no means new. The soaps have been known since the early days of the Roman Empire and are still by far the most widely used substances of this class. The sulphonated oils, produced by the action of concentrated sulphuric acid upon oils of animal or vegetable origin are more than a hundred years old. Up until the beginning of the last war, these two classes repre-

* Present address: Oakite Products, Inc., New York, N. Y.

sented the only surface-active compounds available. But it is not our purpose in the present paper to discuss the soaps and sulphonated oils except in so far as they form the background for the newer synthetics.

The term "surface-active agents" usually refers to the synthetic organic compounds displaying surface-activity, which have been produced so extensively in industry during recent years as wetting-agents, detergents, penetrants, emulsifying-agents, dispersing-agents, foamers and so forth. These compounds have been developed entirely since the last war and particularly during the last fifteen years. The problem presented by the chemistry of surface-active agents is to determine what structural characteristics of the compounds are responsible for their power to promote wetting, detergency, and related surface-active phenomena.

I wish that it were possible to present a quantitative correlation between the chemical constitution of surface-active compounds and their surface-active behavior. However, in the present state of our knowledge it is not possible to do so, since only the meagerest data are available in the literature upon which to base such a correlation. Nevertheless, the subject is well worth considering, not only because of its scientific interest, but because of its great practical value. Perhaps if no more is accomplished than to state the problem clearly in proper relation to its background, a contribution will have been made to the subject.

There are good reasons for the lack of quantitative data on the relation between structure and surface-activity in surface-active agents. In the first place, as these compounds are prepared in industry, they ordinarily consist of complex mixtures of isomers or homologues which are exceedingly difficult to separate into pure components. Such mixtures arise from the nature of the raw materials used, generally fats, which, in turn, are complex mixtures. Quite often, such mixtures function better as surface-active agents than the pure components isolated from them or prepared independently.

Moreover, the physical properties of the mixtures obtained in the preparation of surface-active substances are such as to render their purification quite difficult. They ordinarily possess a gelatinous or soapy consistency and only crystallize with difficulty. Since they generally consist of salts of high molecular weight, they cannot be purified by distillation. Finally, the mixture of organic compounds is contaminated with a considerable quantity of inorganic salt, generally sodium sulphate, the last traces of which are quite troublesome to remove.

Not only do surface-active substances consist of complex mixtures, difficult to purify, but the methods for their evaluation are, of necessity, empirical. This is due to the fact stated previously that such phenomena as wetting, detergency, dispersion and emulsification are exceedingly complex and cannot be adequately estimated in terms of a single easily measurable physical property, such as the ability of the compound to lower the surface tension of water. However, the contrary opinion seems to be quite widespread, as has been pointed out by Williams, Brown and Oakley[1] and by Adam.[2] In fact, the literature is full of so-called wetting and detergency tests based upon indirect measurements such as drop number,[3] suspending power for fine particles,[4] foam number, and other properties.[5]

Such properties of surface-active compounds, while certainly bearing a close relation to the phenomena of wetting, detergency, emulsification and the like, are wholly inadequate as a measure of the effectiveness of any given surface-active compound as a wetting-agent, detergent, or emulsifying-agent. We are, therefore, compelled to resort to direct, empirical methods. These consist of standardized tests approximating as nearly as may be the conditions under which the compound is to be used. Needless to say, such methods leave much to be desired in the way of accuracy and reproducibility.

Up to the present time, surface-active agents have been developed for the most part to meet the needs of the textile industry, hence the methods for their evaluation are closely related to textile processes. For example, the most widely used test for wetting power is the so-called Draves-Clarkson test;[6] in which the time required to wet out a 5-gram skein of raw cotton yarn is measured under specified conditions. This test has been adopted as the official method of the American Association of Textile Chemists and Colorists.

No more adequate measure of detergent efficiency has been evolved than actual washing tests conducted on pieces of fabric soiled with a standard artificial soil, the brightness of the sample being measured photometrically before and after washing. Such tests are described by Rhodes and Brainard,[7] Götte,[8] and by Dreger *et al.*[9]

---

[1] **Williams, E. T., C. B. Brown & H. B. Oakley.** Wetting and Detergency 162 ff Chemical Publishing Co., Brooklyn, N. Y. 1937.
[2] **Adam, N. K.** J. Soc. Dyers Colourists, **53**: 122–129. 1937.
[3] **Kind, W., & J. Auerbach.** Melliand Textilber. **7**: 775–780. 1926.
[4] **Fall, P. H.** J. Phys. Chem. **31**: 801–849. 1927.
[5] **Lederer, E. L.** Kolloidchemie der Seifen, Handbuch der Kolloidwissenschaft in Einzeldarstellung 5. T. Steinkopff, Dresden and Leipzig. 1932.
[6] **Draves, C. Z., & R. G. Clarkson.** American Dyestuff Reptr. **20**: 201–208. 1931: cf. also Yearbook Amer. Assoc. Textile Chemists and Colorists: 199–206. 1944.
[7] **Rhodes, F. H., & S. W. Brainard.** Ind. Eng. Chem. **21**: 60–68. 1929.
[8] **Götte, E.** Kolloid Z. **64**: 222–227. 1933.
[9] **Dreger, E. E., G. I. Keim, C. D. Miles, Leo Shedlovsky & J. Ross.** Ind. Eng. Chem. **36**: 610–617. 1944.

Although surface-active compounds cannot be evaluated for their power to promote wetting, detergency and the like phenomena by the measurement of a single physical property, it should, nevertheless, be recognized that the physical properties of surface active solutions are closely related to such phenomena, and it is possible that the careful measurement of a number of properties such as contact angle, lowering of surface tension, interfacial tension and others, using pure compounds, may permit the establishment of correlations between such measurements which will serve as a means of estimating the more complex phenomena of wetting, detergency and emulsification.

In any event, no adequate treatment of the problem of quantitative correlation between chemical constitution and surface activity in surface active compounds will be possible until vastly more work has been done in the field. This will necessitate the preparation of a large variety of synthetic organic surface active compounds of definite structure and established purity. Such compounds will have to be systematically evaluated either by carefully standardized empirical tests or by such a combination of physical methods as was referred to above.

In the course of these studies, the structures of each class of compounds will need to be systematically varied. Due consideration will have to be given to the effect of such factors as the nature, number and location of hydrophilic groups, the degree of unsaturation of the molecule, the nature and position of side chains, the presence of aromatic or heterocylic rings, the presence of nitrogen or sulphur atoms in the carbon chains, or of ester, amide or ether linkages and other structural features.

II. CHEMICAL CONSTITUTION OF SURFACE ACTIVE AGENTS

Notwithstanding the inadequacy of the existing data for purposes of quantitative correlation between chemical constitution and surface active behavior in surface active agents, the general structural characteristics of surface active molecules are quite well known. For any given compound to display surface activity, there must be present in the molecule a hydrophobic, hydrocarbon portion, to which is attached one or more hydrophilic groups. Many surface active agents belong to the class to which Hartley has given the name, "paraffin chain salts";[10] namely, substances consisting of a straight paraffin chain of

[10] Hartley, G. S. Aqueous Solutions of Paraffin Chain Salts. Hermann and Cie. Paris. 1936.

eight or more carbon atoms, to one end of which is attached an ionic group.

Surface active compounds are classified as anionic or cationic, depending upon whether the hydrocarbon portion of the molecule acquires a negative or a positive charge upon ionization. To these two classes must be added a third, the non-ionic surface active compounds which possess water solubility in virtue of the multiplication of weakly hydrophilic groups in the molecule.

Surface active agents are representatives of the class of compounds known as colloidal electrolytes and are thus able to form ionic micelles in a reversible manner. It is evident that the balance between the hydrophilic and hydrophobic portions of the molecule of a surface active agent is of paramount importance. For example, sodium acetate, which is heavily weighted on the hydrophilic side, is exceedingly water soluble and displays no surface-active properties. The sodium soaps of more than eighteen carbon atoms, on the other hand, are so heavily weighted on the hydrocarbon side as to be too insoluble at ordinary temperatures for effective display of surface activity. Sodium laurate, however, owes its marked surface active properties to just the proper balance between the hydrophilic and hydrophobic groups in the molecule.

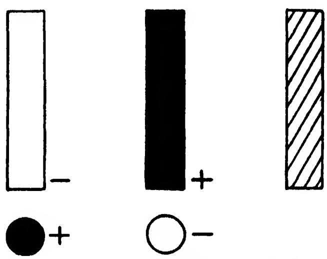


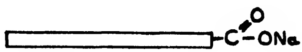
FIGURE 1. Three types of surface active agents.

The types of chemical structure which have been shown to display pronounced surface activity are represented in the figures which follow. No attempt has been made to include all the surface-active compounds reported in the patent and technical literature. Indeed, such a task would require a sizable monograph. Instead, the type formulas have been restricted to those of well-known, widely used commercial surface-active agents, the effectiveness of which has been amply demonstrated.

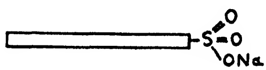
A conventional method of representation has been adopted so as to focus attention upon the characteristic structural features of each type of molecule, while avoiding the distraction which might arise from continued repetition of carbon and hydrogen atoms. Paraffinic carbon chains are represented by rectangular strips, the length of which is proportional (except in FIGURE 3) to the actual length of the carbon chain present in the compound. The presence of olefinic bonds is indicated by two heavy parallel lines. Other structural features are represented in accordance with the usual conventions of organic chemistry.

It is possible only to a very limited extent to assign a characteristic kind of surface activity to any given structure. This is due to the

I



II



III

IV

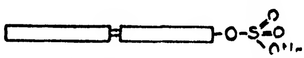


FIGURE 2 I Sodium stearate II Cetyl sodium sulphonate. III. Sodium sulphate of lauryl alcohol. IV. Sodium sulphate of oleyl alcohol

fact, stated previously, that most surface active agents possess all of the so-called properties of wetting, detergency, emulsifying-power and the like to some degree. It is quite rare for a given surface-active compound to display only one property such as wetting to the exclusion of the remainder.

1. Anion-active compounds.

(a) *Paraffin chain salt types* (FIGURE 2). This class embraces the ordinary soaps, the sulphated fatty alcohols and the true alkyl sulphonates. These compounds find their greatest use as detergents, although the fatty alcohol sulphates, in particular, are widely used as wetting agents. However, they are distinctly inferior for this purpose by comparison with the powerful wetting agents to be described later.

(b) *Alkylated aromatic sulphonates* (FIGURE 3). These compounds were originally developed as soap substitutes[11] and have enjoyed recognition for the most part as detergents. However, with the proper re-

[11] Price, Donald. Am. Ink Maker 22 (6) 21–24, 45. 1944.

lationship between chain length and molecular weight, as well as the proper position of the side chains, compounds of this class may be made to function effectively as emulsifying- and wetting-agents

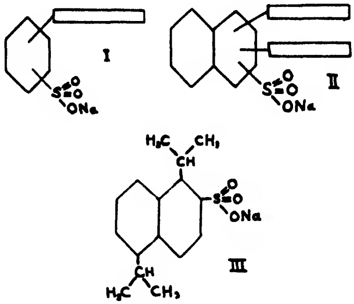


FIGURE 3 I Mono alkyl benzene sodium sulphonate II Di alkyl naphthalene sodium sulphonate III Di isopropyl naphthalene sodium sulphonate (Nekal)

(c) *Straight chain compounds,* in which the hydrophilic group is more complex than in the case of the simple paraffin chain salt types (FIGURE 4)

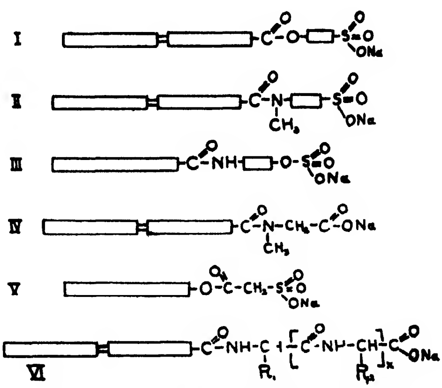


FIGURE 4 I Sodium salt of ester of oleic acid and isethionic acids (Igepon A) II Sodium salt of amide of oleic acid and methyl taurine (Igepon T) III Amide of lauric acid and sulphated monoethanolamine IV Sodium salt of oleyl amide of N-methyl glycine V Sodium salt of lauric ester of sulpho-acetic acid VI Condensation product of oleic acid with degraded protein

Just as the fatty alcohol sulphates were developed in an effort to overcome the sensitivity of the soaps to hard water and acids, the same end is accomplished in the present class of compounds by blocking the carboxyl group by the formation ester or amide linkages, while, at the

same time, introducing a less sensitive hydrophilic group. The structures vary so widely and their details play so great a part in the surface active behavior of the resulting compounds that it is scarcely possible to assign any characteristic type of behavior to this class.

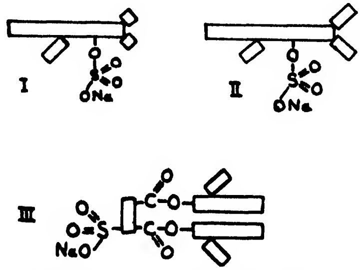


FIGURE 5. I. Sodium sulphate of 2-methyl-7-ethyl-undecanol-4. II. Sodium sulphate of 3,9-diethyltridecanol-6. III. Di-(2-ethyl-butyl) sodium sulphosuccinate.

A large part of the patent literature covering surface-active agents is devoted to compounds belonging in the present category. The trend has been in the direction of greater and greater complexity of structure in the compounds described, no doubt, for the purpose of circumvent-

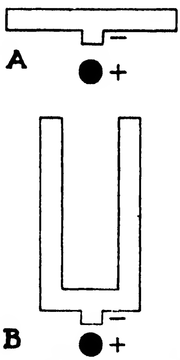


FIGURE 6. Hypothetical shapes of surface active agents with hydrophilic group in center of molecule.

ing prior patents. However, the alleged superiority of many of these complex substances is somewhat doubtful.

(d) *Compounds in which the hydrophilic group is located near the middle of the carbon chain rather than at the end* (FIGURE 5).

The compounds belonging to this class may have one of several configurations, as shown in FIGURE 6. They are the most powerful wetting-

agents and penetrants known. In fact, one of the few generalizations that can be made with regard to the relation between chemical constitution and surface activity is that the location of a hydrophilic group in the middle of a hydrocarbon chain strongly favors wetting-power at the expense of other surface active properties.[12] Support for this generalization is to be found in recent work presented by Dr. Leo Shedlovsky elsewhere in the present symposium.[9]

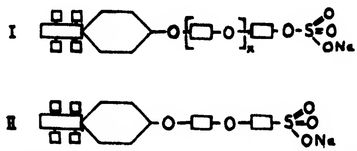


FIGURE 7 I Sodium sulphate of polyethylene ether of 1,1,3,3-tetramethylbutyl phenol. II. Sodium sulphonate of polyethylene ether of 1,1,3,3-tetramethylbutyl phenol

(e) *Compounds containing both aromatic rings and complex carbon chains* (FIGURE 7).

This class of compounds is relatively new, but undoubtedly possesses both wetting and detergent properties to a rather marked degree.

(f) *Heterocyclic compounds* (FIGURE 8).

Comparatively few heterocyclic compounds have been developed for use as surface-active agents, perhaps because of the factor of cost.

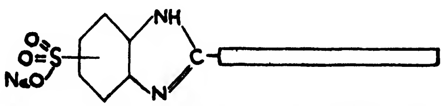


FIGURE 8 Sodium sulphonate of heptadecylbenzimidazole.

Compounds of the type shown in FIGURE 8 are said to be used chiefly as lime-dispersing agents.

2. Cation active compounds (FIGURE 9)

Although the properties of cation-active surface active compounds are just beginning to be thoroughly understood, compounds of this class were among the earliest to be developed.[13] In contrast to many of the anion-active types, they are quite stable and extremely effective in acid solution. They have found wide use in the textile industry as finishing agents, but also possess foaming, wetting and detergent properties as was pointed out as early as 1913 by Reychler.[14]

[12] Wilkes, B. G., & J. N. Wickert. Ind. Eng. Chem. 29: 1234–1239. 1937.
[13] Hartmann, M., & H. Kagi. Zeit. angew. Chem. 41: 127–130. 1928.
[14] Reychler, A. Bull. soc. chim. Belg. 27: 217–225. 1913.

Great interest has been aroused in this class of compounds as a consequence of the discovery by Domagk[15] of the powerful bactericidal properties of the higher alkyl quaternary ammonium salts. This subject is discussed in detail by Dr. Valko elsewhere in the present symposium [15a]

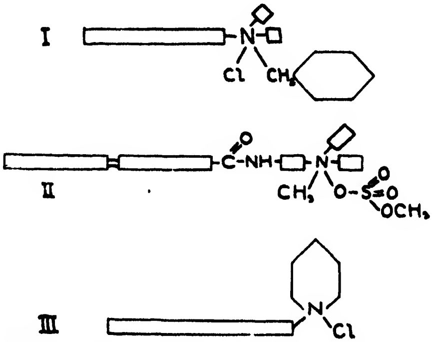


FIGURE 9. Di-methyl-lauryl-benzyl ammonium chloride. II. Methosulphate of oleyl amide of diethylethylenediamine III. Cetyl pyridinium chloride.

3. Non-ionic compounds (FIGURE 10)

This is the newest class of surface-active agents to be developed. Although their surface active properties do not appear to be quite as powerful as those of the ionic compounds, they enjoy certain advantages over the latter, due to the absence of ionic groups. For example,

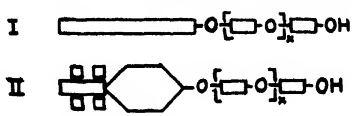


FIGURE 10 I. Polyethylene ether of lauryl alcohol. II. Polyethylene ether of 1,1,3,3-tetramethylbutyl phenol.

they are effective under conditions of hard water and pH where many ionic surface-active agents would be precipitated, or otherwise rendered ineffective.

Besides the types of non-ionic surface-active agents shown in the figure, several others have recently come into the field, among which may be mentioned mannitan and sorbitan esters of higher fatty acids and polyethylene oxide derivatives thereof. Compounds of the latter class have recently been reviewed by Goldsmith.[16]

15 Domagk, G. Deut. med. Wochenschr. 61: 829–832. 1935.
15a Valko, E. I. Ann. N. Y. Acad. Sci. 46 (6): 471–478. 1946.
16 Goldsmith, H. A., Chem. Rev. 33: 257–349 1943.

III. RELATIONSHIP BETWEEN CONSTITUTION AND PERFORMANCE OF SURFACE ACTIVE AGENTS

As was pointed out earlier, far more work will have to be done in the direction of the synthesis of surface-active compounds of definite structure and purity, together with a study of their surface active properties, before any extensive correlation between chemical constitution and surface activity will be possible. Although the existing data are far too meagre to form the basis for such a correlation, some valuable work of the kind has been done which merits consideration. The papers to be discussed in what follows constitute a step in the right direction, which it is to be hoped will be followed up by other investigators in the future.

Götte[17] studied the detergent power of a homologous series of saturated alkyl sodium sulphates C_{12}, C_{14}, C_{16}, and C_{18}. The compounds were prepared by reduction of the butyl esters of the corresponding fatty acids with sodium and alcohol followed by sulphation of the resulting alcohols. Purification was effected by repeated vacuum distillation of the alcohols so as to provide pure raw material for the sulphation step and by repeated crystallization from alcohol of the sodium sulphates in order to remove inorganic salts.

The detergent power of the compounds was evaluated over a pH range from 1 to 13 by a procedure devised by the author. Normal bleached cotton was treated with colloidal carbon and a mixture of mineral and vegetable oils. The fabric was then cut into squares and washed by tumbling in glass jars rotated in a constant temperature bath at 60° C. Each jar contained one liter of solution at a concentration of one gram per liter of detergent. The solutions were buffered with 200 c.c. of various buffer mixtures and the author claims that the amount of salts added was insufficient to have a disturbing influence.

The brightness of the samples was measured with a Pulfrich step photometer before treatment with the artificial soil and after washing. The results were expressed as per cent brightness based on that of the original untreated fabric taken as 100. When per cent brightness was plotted on a logarithmic scale as a function of pH, curves were obtained showing maxima at a pH slightly above 10 and minima in the region 4–5.

The author also derives a quantity Q, to which he gives the name, detergent value, as follows:

[17] Götte, E. Kolloid Z 64: 222–227; 327–331; 331–335 1933

$$\text{Log. } W_m - \text{Log. } W_o = \text{Log. } \frac{W_m}{W_o} = \text{Log. } q = Q.$$

Where W_m = per cent brightness of sample after washing with solution containing detergent.

W_o = per cent brightness after washing with water alone (plus buffer salts)

The values of Q were then plotted as a function of pH. It is obvious that the value of W_m will be greater than W_o for any detergent which is more effective in removing soil than plain water. As a consequence, Q will have positive values. Where Q has negative values, the detergent is less effective than water and tends to redeposit the dirt upon the fabric.

At the temperature used in his experiments, 60° C, Götte found the detergent power of the sodium alkyl sulphates to lie in the order $C_{16} > C_{14} > C_{18} > C_{12}$. Measurements of foam number by the Stiepel method[18] showed the compounds to lie in the same order at 60° C. But at 40° C the C_{14} compound produced the maximum foam, while at 20° C the C_{12} compound was the most powerful foamer.

One cation-active compound, tricthyl-alkyl ammonium chloride (where alkyl denotes the carbon chains derived from coconut oil fatty acids) was evaluated by Götte's procedure. The values of Q for this compound were found to be negative for all pH values, so that the entire curve lay below the x-axis. Götte interpreted this to mean that compounds of this type act to redeposit dirt upon the fabric rather than to function as detergents. However, as pointed out by Valko,[19] this conclusion of Götte's is not in agreement with earlier findings of Reychler.

Venkataraman and co-workers[20, 21, 22] prepared a long series of fatty amides of aromatic amino sulphonic acids and studied the wetting power and certain other properties of the compounds. The structures were varied so as to determine the effect of such factors as the position of hydrophilic groups, unsaturation in the carbon chains, molecular weight and the presence of halogen or methyl groups.

The method of preparation, which, in general, consisted of condensing the appropriate amino sulphonic acid with the required fatty acid chloride in the presence of pyridine, is described for each compound. The methods of purification used for many of the compounds are also

[18] Stiepel, C. Seifensieder Ztg. 41: 347. 1914.
[19] Valko, E. "Kolloidchemische Grundlagen der Textilveredlung". 635–636. J. Springer, Berlin. 1937.
[20] Dhingra, D. R., J. S. Uppal & K. Venkataraman. J. Soc. Dyers Colourists 53: 91–100. 1937.
[21] Uppal, J. S., & K. Venkataraman. Ibid. 55: 125–134. 1939.
[22] Shirolkar, G. V., & K. Venkataraman. Ibid. 57: 41–49. 1941.

given as well as analytical figures based upon the determination of Na and S.

The wetting power of the compounds was determined both by an improved Herbig procedure developed by the authors[23] and by the Draves test. Since the Herbig method is unfamiliar to American readers, while the Draves test is quite well-known, only values obtained by the latter method are presented here. It is noteworthy that the order in which the compounds fall does not entirely coincide in the two methods. This disagreement illustrates the difficulties inherent in estimating such complex phenomena as wetting and detergency.

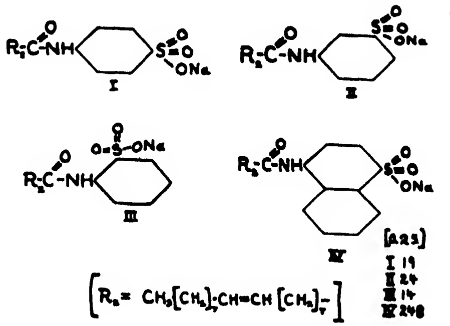


FIGURE 11. Figures at lower right denote sinking time in seconds in Draves test.

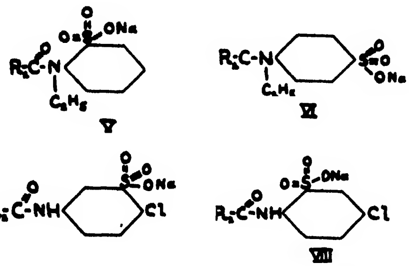
It is to be regretted that the authors, although they prepared some fifty compounds in all, did not vary the structures more systematically, carrying each variation through a complete series, so as to observe the effect of a single variable with all other factors constant. As it is, their data are insufficient to support any general conclusions, although the effects of certain variations in structure are clearly evident.

Compounds selected from several of Venkataraman's papers are shown in the following figures and illustrate the effect upon wetting power of certain structural changes. The sinking times in seconds in the Draves test, at a concentration of 0.25% of wetting agent, are given in the lower right hand corner of each figure.

In FIGURE 11, are shown the sodium sulphonates of the amides obtained by condensing orthanilic, metanilic and sulphanilic acids, respectively, with oleic acid. Although the differences are not of a high order, the effect of the position of the hydrophilic group is quite evi-

[23] Forster, R. B., J. S. Uppal & K. Venkataraman. Ibid. 54: 465–472. 1938.

dent, the ortho sulphonate being the most effective wetting agent. The effect of increased molecular weight in this series is shown by the substitution of a naphthalene for a benzene nucleus. Compound IV is seen to be far inferior to Compound I.



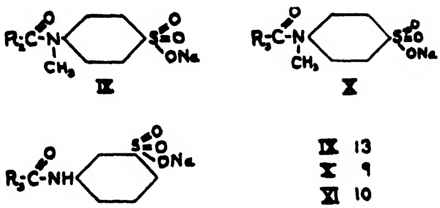
V VI VII VIII

[$R_2 = CH_3 \cdot [CH_2]_7 \cdot CH{=}CH \cdot [CH_2]_7 \cdot CH_2-$]

V	19
VI	100
VII	50
VIII	21

FIGURE 12. Figures at lower right denote sinking time in seconds in Draves test

FIGURE 12 shows the effect of varying the position of the hydrophilic group in two other series of compounds. Comparing Compound V with Compound VI, and VIII with VII, the ortho position of the hydrophilic group will again be seen to give the most effective wetting agent. This time the effect is more striking than in the previous case.



IX X XI

IX	13
X	9
XI	10

[R_3 = Mixed Coco Chains]

FIGURE 13. Figures at lower right denote sinking time in seconds in Draves test.

FIGURE 13 shows the effect of a change in the nature of the fatty acid chain. R_3 in this figure represents the mixed fatty acid chains derived from coconut oil. By comparing Compound IX with Compound X, and XI with I (FIGURE 11), this variation will be seen to have a favorable effect upon the wetting power of the compounds.

Caryl[24] reported the wetting power and solubility of 36 esters of sodium sulphosuccinic acid,

$$\begin{array}{l} \quad\;\; CH_2\text{–}COO\text{–}(CH_2)_n\text{–}CH_3 \\ \quad\;\; | \\ NaSO_3CH\text{–}COO\text{–}(CH_2)_n\text{–}CH_3 \end{array}$$

Wetting power was measured by the Draves test and his results are expressed in terms of the grams per liter of wetting agent required to give a 25-second sinking time by the Draves test at 30° C.

Diesters, mixed esters and salts of mono-esters were included in Caryl's paper. The compounds were not especially purified, but the author stated that their purity was of the order of 99%. The seven best wetting agents of the series are given in TABLE 1.

TABLE 1

Ester	Gr./liter to give 25 sec. sinking time.
1. Di (1-methyl-4-ethyl hexyl)	0.16
2 Mono-2-ethyl hexyl mono-1-methyl heptyl	0 16
3. Mono-2-ethyl hexyl Mono-1-methyl-4-ethyl hexyl	0.17
4. Di(1-methyl heptyl)	0.19
5. Di(2-ethyl hexyl)	0.20
6. Di(1-isobutyl-3-methyl butyl)	0.21
7. Di(1-butyl amyl)	0.22

As previously stated and as will be readily seen from the data, compounds of this general structure are among the most powerful wetting agents known. It is noteworthy that esters of the corresponding tribasic acid (sodium sulphotricarballylic acid) are of the same order of wetting power.[25]

Neville and Jeanson[26] studied the relation between structure and surface tension in a series of mono- and di-alkyl benzene sodium sulphonates. The compounds included in their studies were the sodium sulphonates of benzene, toluene, ethyl benzene, isopropyl benzene, butyl benzene, xylene and cymene.

Ethyl benzene, isopropyl benzene and butyl benzene were especially prepared by the Fittig synthesis and all of the compounds were sul-

[24] **Caryl, C. R.** Ind. Eng. Chem **33:** 731–737. 1941

[25] **Price, D.** Brit. Patent **551,246** (to National Oil Products Co.); **Nawiasky, P., & G. E. Sprenger.** U. S. Pat **2,315,375** (to General Aniline and Film Corp.).

[26] **Neville, Harvey A., & Chas. A. Jeanson, III.** J. Phys Chem. **37:** 1000–1008. 1933

phonated by the method of Gattermann.[27] The sodium sulphonates were isolated either by precipitation with sodium chloride or by evaporation to dryness, and then purified by crystallization from absolute alcohol or acetone until salt free. With the exception of benzene sodium sulphonate they undoubtedly consisted of mixtures of isomers.

The surface tension measurements were carried out over a range of concentrations from 0 to 0.5 moles per liter at 18° C. with an improved torsion balance described by de Gray[28] and surface tension plotted as a function of concentration.

In the case of benzene sodium sulphonate, the observed lowering of surface tension was roughly proportional to concentration, resembling in this respect the lower aliphatic acids and, in general, substances in true solution which lower surface tension. Beyond the first member, the curves showed an increasing tendency to sag or deviate from the straight line relationship.

In the cases of the pairs of compounds having identical molecular weight, ethyl benzene sodium sulphonate and xylene sodium sulphonate, that with two substituents was considerably more effective than that with one. The same was true for the pair, butyl benzene sodium sulphonate and cymene sodium sulphonate. The curves for isopropyl benzene sodium sulphonate, cymene sodium sulphonate and butyl benzene sodium sulphonate exhibit minima characteristic of active surface tension depressants, indicating a colloidal condition of the solute.

In the sulphonation of toluene, a mixture of two isomers is obtained consisting of about 70% para and 30% ortho. These isomers were separated by taking advantage of the difference in solubility of their sulphonyl chlorides, and surface tension measurements were made on solutions of the pure isomeric sodium sulphonates. Ortho toluene sodium sulphonate was found to be considerably more effective than the para isomer, while the mixture lay intermediate between the two. At lower concentrations, the activity of the ortho isomer predominated while, at higher concentrations, the surface activity of the mixture approached that of the less active para compound.

Davis, Price and Milligan[29] prepared a series of sulphonated amides of the general structure,

$$\begin{array}{ccccccccc} & & & R & & CH_3 & & \\ & & & | & & | & & \\ CH_3(CH_2)_n & - & CO - & N & -CH_2- & C & -CH_2SO_3Na \\ & & & & & | & & \\ & & & & & OH & & \end{array}$$

[27] ———, — Practical Methods in Organic Chemistry: 280. Eng. Ed. 1915.

[28] De Gray, R. J. Ind. Eng. Chem., Anal. Ed. 5: 70–73. 1933.

[29] Davis, G. U., D. Price & J. G. Milligan. U. S. Pat. 2,367,010 (to National Oil Products Co.).

in which both the fatty acid chain and the side chain attached to the nitrogen were varied. The wetting power of these compounds was estimated by the Draves test,[30] the results of which are shown in TABLES 2 and 3. The products were not crystalline, but possessed a waxy or gelatinous consistency and, while probably not analytically pure, were of the order of 98–99% purity.

TABLE 2

R	No. of Carbons in R	Acid	Draves sinking times			
			.2%	.1%	.05%	02%
Hydrogen	0	Lauric (12)	37	103		
Methyl	1	Lauric (12)	15	35		
Isopropyl	3	Lauric (12)	9	17		
Butyl	4	Lauric (12)	6	9	20	88
Amyl	5	Lauric (12)	6	11	20	274
2-Ethyl butyl	6	Lauric (12)		10	18	
2-Ethyl hexyl	8	Lauric (12)		13	29	
Phenyl	6	Lauric (12)	18	28	30	

It will be seen from TABLE 2 that lengthening the nitrogen side chain has a profound effect upon the wetting power of the compounds. A maximum is reached when R = butyl, after which wetting power begins to drop off. In like manner, the length of the fatty acid chain markedly affects the wetting power of the compounds as shown in TABLE 3. In this case, the compound derived from lauric acid is the most efficient wetting agent of the series.

Although the data are insufficient to justify any general conclusions, it would appear that the sum of the number of carbon atoms in the fatty acid and in the nitrogen side chain is the critical factor in determining wetting power in this series. The indications are that the optimum value for this sum is about sixteen. In other words, a compound derived from capric acid having a n-hexyl side chain attached to the nitrogen would probably be as effective as one derived from lauric acid with a butyl side chain.

By far the most outstanding contribution to the study of the relation between chemical constitution and surface activity up to the present time is that of Dreger *et al*, covering a series of sodium alcohol sulphates the structures of which were systematically varied. These authors took the pains to prepare pure compounds of definite

[30] Draves, C. Z., & R. G. Clarkson. American Dyestuff Reptr. 20: 201–208 1931.

TABLE 3

R	Acid	No. of carbons in acid	.2%	.1%	.05%	.02%
Butyl	Caprylic	8	146			
Butyl	Capric	10	6	37	420	
Butyl	Lauric	12	6	9	20	88
Butyl	Myristic	14	36	39	95	
Butyl	Palmitic	16	109	201		
Butyl	Coconut fatty acids		14	32		
Phenyl	Capric	10		15	48	
Phenyl	Lauric	12	18	28	30	
Phenyl	Coconut F.A.			56	171	
2-Ethyl hexyl	Caprylic	8	4	10	55	
2-Ethyl hexyl	Capric	10	3	8	15	
2-Ethyl hexyl	Lauric	12		13	29	
2-Ethyl butyl	Capric	10		7	20	
2-Ethyl butyl	Lauric	12		10	18	

structure, for which they give complete analytical figures. Moreover, their measurements of the surface active properties of the compounds were carried out by means of empirical tests which were carefully standardized by themselves.

Further reference to this work is omitted here, as it is discussed in detail by Dr. Leo Shedlovsky elsewhere in the present symposium.[31] It should be noted, however, that these authors have set a standard in the study of chemical constitution in relation to surface activity which will have to be followed by later investigators, if their work is to merit serious consideration.

DISCUSSION OF THE PAPER

Dr. E. I. Valko (*Onyx Oil and Chemical Co., Jersey City, N. J.*):

An interesting attempt to correlate chemical structure and surface properties of paraffin-chain salts was made a few years ago by Hartley.[1] Hartley pointed out that, after the critical concentration of the micelle formation is reached, the reduction of the interfacial tension between water and a non-polar liquid remains at a constant value when the concentration of the surface active agent is further increased. The reason for this behavior is that the micelles, being symmetrically hydrophilic, are not surface active and, consequently, can not participate in the reduction of the interfacial tension. If micelle formation could be prevented, a greater reduction of the interfacial tension with increasing concentration would be expected. Substitution of the straight chains by branched or double chains is a means to prevent micelle formation. Hartley confirmed this assumption by preparing the sulfonates of the di-alkyl ethers of dihydric phenols as well as the sulfonates of the alkyl ethers of phenol and comparing their interfacial activity. At low concentration, the sulfate of hexadecyl ether of phenol showed a higher

[31] Shedlovsky, Leo. Ann. N. Y. Acad. Sci. 46 (6): 434–439. 1946.
[1] Hartley, G. S. Trans. Faraday Soc. 37: 130. 1941.

interfacial tension than the sulfonate of the di-octyl ether of resorcinol. With increasing concentration, the hexadecyl compound reached a constant value of the interfacial tension, while the di-octyl compound continued to reduce the interfacial tension far below the minimum value exhibited by the hexadecyl compound. A disadvantage of the double chain or branched chain compounds is their comparatively low solubility which is also due to the lack of micelle formation.

Dr. Foster Dee Snell (*Brooklyn, New York*):

Dr. Price has given us a very interesting summary. As one small emendation, because he has been speaking of agents usually listed as surface active, he has missed one large tonnage item. I refer to fatty monoglycerides made usually by alcoholysis of triglycerides with glycerine catalyzed by soap. They are used mainly in shortenings and margarines in quantities ranging from several per cent down to a half per cent. A minor use is in cosmetics In terms of actual tonnage used, they are among the first five

We consider detergency to be a compound property built up from two factors It has been our experience that detergents must pass stringent tests on interfacial tension and on dispersing power. The latter is, in turn, an inseparable composite of deflocculating power and emulsifying power for oil-coated solid particles. In line with remarks by Dr. Fischer on the Draves Test and by Dr. Price on Launderometer tests, we believe, that a careful study of those two properties, properly interpreted and followed by a few practical qualitative tests, is more satisfactory than performance tests which are designed to incorporate all of the factors which go into some form of a commercial use. And, incidentally, the latter factor, dispersing power, is usually controlling. It is our experience that we can pretty well predict from structure how a surface-active agent will perform as a detergent, and that subsequent experiment confirms the prediction in at least 90% of the cases

Lastly, to confound confusion, I want to mention a peculiar case of surface activity. Carbowax, which is approximately 44 molecules of ethylene oxide polymerized together, has negligible surface activity in water solution, but is an excellent surface-active agent in 30% sodium bisulfate solution, and will reduce the surface tension by about 30 dynes I use that illustration because the molecule has no long non-polar group In terms Dr Price has used, the structure is interrupted by an oxygen atom after each two carbon atoms. One can figure out how it orients to be effective in a solution where its solubility has been sufficiently reduced

Mr. L. H. Flett (*National Aniline Division, Allied Chemical and Dye Corporation, New York, N Y.*)

Dr Price has confined his remarks to the commercially interesting surface active agents For this reason, he has confined his remarks to products all of which have a long aliphatic group. Such products are commercially interesting for two reasons:

(1) Simple organic compounds with a long saturated or substantially saturated aliphatic chain are generally colorless. Industry, in general, prefers to work with colorless surface active agents.

(2) Products with long aliphatic chains are generally derived from oils or petroleum which are cheap sources of raw material.

In spite of the commercial interest in products with long aliphatic chains described by the speaker, it must be remembered that surface active agents do not require such an aliphatic chain. Surface active agents may be found in every class of the organic compounds. For example, strictly aromatic compounds or cyclo-aliphatic compounds may be surface active agents.

The speaker has brought out in a very excellent way the inability of industry to classify products by any simple test. It would be desirable if a simple test could be developed by which all surface active agents could be classified. The best one could then be chosen and used for every purpose. This very definitely cannot be done. Different surface active agents are found particularly effective for different

uses. One product may be good for washing, another may be good as an emulsifying agent and another may be a particularly effective agent for wetting a particular kind of fabric.

Even in the case of one particular use, the relative effectiveness of the different surface active agents will vary under different conditions. For example, in emulsification, it is possible for one surface active agent to be better than another for emulsifying mineral oil and inferior to the other for emulsifying capryl alcohol.

It is expected that time will increase rather than reduce the number of surface active agents.

I like to compare surface active agents with dyestuffs. Each dyestuff has its own particular properties which make it valuable and necessary to industry.

PROPERTIES INVOLVING SURFACE ACTIVITY OF SOLUTIONS OF PARAFFIN CHAIN SALTS

By Leo Shedlovsky

From Colgate-Palmolive-Peet Co., Jersey City, N. J.

Surface active properties of compounds show characteristic changes when their structures and molecular weights are altered. Only a few reports deal with relatively pure compounds. In order to attribute surface properties to the materials used, it is often necessary to avoid even small amounts of certain impurities.

A comparative study of a number of properties involving the surface activity of solutions of paraffin chain salts has been reported recently.[1–4]

In this work, an attempt has been made to avoid impurities as well as complicating variations in molecular structure which may alter the characteristics.

This work may be considered in three parts:

(1) The preparation and properties are described for alternate members of a homologous and an isomeric series of purified sodium salts of secondary alcohol sulfates containing from 11 to 19 carbon atoms and for a straight hydrocarbon chain with the sulfate group in various positions, as well as of the sodium salts of the primary alcohol sulfates with 10, 12, 14, and 16 carbon atoms. The surface tension, interfacial tension (benzene/water), foaming, wetting and detersive properties of solutions of these compounds are reported and discussed.

(2) Typical examples are shown where minima in surface tension-concentration curves are indicated only when certain impurities are present.[2, 3]

(3) Some relative surface active properties of solutions of soaps of pure fatty acids are considered on the basis of their surface and interfacial tension and the foam stability of their solutions.[5, 6, 4]

---

[1] **Dreger, E. E., G. I. Keim, G. D. Miles, L. Shedlovsky & J. Ross.** Ind Eng. Chem. **36:** 610. 1944.
[2] **Miles, G. D., & L. Shedlovsky.** J. Phys. Chem. **48:** 57. 1944.
[3] **Miles, G. D.** J. Phys. Chem. **49:** 71. 1945.
[4] **Miles, G. D., & J. Ross.** J. Phys. Chem. **48:** 280. 1944.
[5] **Powney, J.** Trans. Faraday Soc. **31:** 1510. 1935.
[6] **Powney, J., & C. C. Addison.** Trans. Faraday Soc. **33:** 356, 372. 1937.

I. PREPARATION OF SODIUM ALCOHOL SULFATES

1. Secondary Alcohols from Ketones

The fatty acids or their methyl esters used in the synthesis of the ketones were obtained mainly from natural sources. To eliminate unsaturated acids, present as impurities, the methyl esters of the fatty acids were treated with an excess of potassium permanganate in boiling acetone solution, filtered, washed with a solution of alkali and with water, dried, and crystallized from the acetone solution. These esters were then fractionally distilled.

The ketones were prepared by passing the vapors of the above pure acids or their methyl esters, at 375–400° C., over thorium oxide which served as a catalyst.[7]

The symmetrical ketones were obtained by passing the single fatty acid or ester over the catalyst. The unsymmetrical ketones were made by passing the appropriate mixture of two pure fatty acids or esters over the catalyst. The resulting three ketones, usually with widely different boiling points, were separated by fractional distillation.

The secondary alcohols were made by catalytic reduction of the pure ketones, using a catalyst of nickel and kieselguhr, at 90-120° C and approximately 500–700 pounds per square inch pressure.

2. Secondary Alcohol Sulfates

In general, the method of sulfation was as follows: To 100 grams of acetic acid, 60 grams of chlorosulfonic acid (0.515 gram mole) were added with stirring, keeping the mixture in an ice bath. Then 100 grams (0.5 gram mole) of the secondary alcohol to be sulfated were added gradually during five minutes, and the whole mixture was stirred for another 30 minutes at 4° C.

The reaction mixture was poured on 300 grams of cracked ice, 300 cc. of n-butanol were added, and the solution was neutralized with 2 N sodium carbonate solution and sufficient solid sodium bicarbonate to keep the solution saturated with inorganic sodium salts. The neutral sodium alcohol sulfate was separated with the butanol layer, and the aqueous layer was further extracted with four successive portions of butanol. By concentrating the combined butanol extract under vacuum, water was removed and the precipitated solid inorganic salts were separated by filtration. The remaining butanol was then removed by distillation at a pressure of 40 mm. of mercury, and, toward the end of the distillation, 2 liters of distilled water were added and

[7] Organic Syntheses 16: 47. 1936.

distillation was continued until all the butanol was removed. The pH of the solution was adjusted to 7.0. The cold aqueous solution was then extracted with ethyl ether to remove unsulfated material Finally, the ether in the aqueous solution was removed by concentrating under vacuum. In this way, a solution of the sodium alcohol sulfate was obtained, whose concentration was determined by evaporating a portion to dryness at 80° C. in a vacuum oven. When possible, the sodium alcohol sulfate was crystallized from the dry butanol solution and then recrystallized from distilled water.

The yields of sodium sec-alcohol sulfate were generally about 80–95% of the theoretical amounts. The sodium sulfates of the methyl alkyl sec-carbinols were crystalline solids The salts of the other secondary alcohol sulfates were obtained only in aqueous solution, and they showed decreasing solubility with increase in molecular weight and the higher members of the series, sodium pentadecane-8-sulfate (15–8), sodium heptadecane-9-sulfate (17–9), and sodium nonadecane-10-sulfate (19-10) formed gelatinous precipitates. (The first figure in parentheses following these compounds refers to the number of carbon atoms, and the second number, to the position of the sulfate group. In order to avoid cumbersome repetition, this convention will be used hereafter.)

3. Primary Alcohol Sulfates

These sulfates were prepared from the corresponding purified primary alcohols by sulfation according to the procedure described for the sulfation of secondary alcohols.

4. Solutions

The solutions were made with distilled water. The composition of the stock solutions was calculated on the basis of chemical analyses The concentration of the stock solution of sodium pentadecane-4-sulfate (15-4) was also checked by the Karl Fischer reagent[8] on a weighed sample which had most of the water evaporated from it.

II. PREPARATION OF SOAPS

1. Fatty Acids

The fatty acids used in preparing the soaps were derived from natural fats and were purified by different methods, according to the nature of the fatty acid. The saturated fatty acids were converted into

[8] Fischer, K. Angew Chem 48: 394. 1935.

the methyl esters and these were treated with potassium permanganate in acetone solution in order to remove the unsaturated impurities. The recovered neutral methyl esters were then fractionally distilled through a 5-foot modified Fenske fractionating column.

The oleic acid was purified through fractional crystallization of the lithium salt from 80 per cent alcohol; the elaidic acid by distillation and fractional crystallization of the acid prepared from purified oleic acid by action of selenium; the ricinoleic acid by fractional distillation of the methyl ester made from castor oil; the undecylenic acid by fractional distillation of the methyl ester of the crude acid prepared from castor oil. The purity of these acids was checked by the acid, saponification, and iodine values and by the melting point.

2. Soaps

The free fatty acids were obtained from the esters by saponification, acidifying, washing, and drying. The soaps were prepared from the fatty acids by neutralizing exactly in alcohol solution, using phenolphthalein as an external indicator. They were obtained in the solid form (0.5–2.0 per cent moisture) by drying on a laboratory drum dryer. In this way, soaps which were substantially neutral and contained only traces ($<$ 0.2 per cent) of free alkali were obtained.

Powney and Addison[5, 6] prepared the soaps by alcohol saponification from fatty acids which were 95% pure. Special precautions were taken to keep the free alkali down to levels of less than 0.06 per cent.

III. DETERMINATION OF PROPERTIES

1. Measurement of Surface Tension and Interfacial Tension

The static surface tension and interfacial tension of the solutions were determined by the du Nouy ring method, employing a calibration curve of the dial readings plotted against the surface tension of pure liquids.[9]

For temperatures above 30° C., a glass cylinder 2½ inches in diameter and 2 inches long was wrapped with asbestos-covered Nichrome wire. A cork stopper was placed in the bottom of the cylinder, and the top had a blackened cover provided with an opening large enough for the wire supporting the ring. The Nichrome wire heater was operated on the 110-volt circuit through an auxiliary variable resistance which was adjusted to obtain the desired temperature. The blackened cover was heated from the outside by a 50-watt bulb placed about 18 inches

[9] International Critical Tables 4: 446. 1928.

above it. This arrangement avoided condensation of water on the sides of the heating compartment. The solutions were warmed to the proper temperature before being placed in the compartment. The containers for this purpose were weighing dishes, 1⅞ inches in diameter and 1⅛ inches high.

In every case, the temperature noted is that of the solution in bulk, ± 0.5° C. Because the temperature coefficient of surface tension for the solutions tested is not large, a variation of 0.5° C. was not considered significant.

The method of calibration tested by Macy[10] avoids the necessity of using the Harkins and Jordan correction factor[11] and gives a precision of about ± 0.3 dyne per cm., but the reproducibility is better than this or about 0.2 dyne per cm. Such a calibration corresponds closely to the surface tension obtained by applying the Harkins and Jordan correction.

In the study of minima in surface tension-concentration curves,[2] particular care was exercised in preparing the solutions and measuring the surface tension, in order to avoid accidental contamination. From the point of view of speed and ease of maintaining clean surfaces, the best procedure of those used was as follows: A du Nouy tensiometer was placed upon a flat plate on top of a screw-jack. This permitted the uniform elevation of the instrument with respect to the surface of the solution. The material to be tested was weighed and placed in a 1000-ml. Erlenmeyer flask which had been carefully cleaned, rinsed with distilled water, and paraffined inside and out around the neck. The solutions were diluted by stepwise addition of measured volumes of distilled water from an automatic buret of 100-ml. capacity to the flask. An extension of glass rod approximately 1 mm. in diameter connected the platinum-iridium ring with the torque arm of the tensiometer so that the ring rested upon the surface of the solutions inside the flask. The neck of the flask was covered with a slotted sheet of suitable material and the measurement was made by gradually pulling the ring away from the surface of the solutions. By the addition of more water, the surface-tension curve was obtained for a concentration range of ten to one before it was necessary to weigh a fresh sample. This method gave a precision of about ± 0.15 dyne per centimeter.

Powney[5] measured essentially static surface tensions of soap solutions using the du Nouy ring method with the solutions at 20° C.

Powney and Addison[6] used a modified drop weight method for the

---

[10] **Macy, R.** J. Chem. Ed. **12**: 573. 1935.
[11] **Harkins, W. D., & A. F. Jordan.** J. Amer. Chem. Soc. **52**: 1751. 1930.

determination of interfacial tension of soap solutions. They point out that the change in the interfacial tension with time is due to two distinct causes, first, the normal diffusion time to the interface, which is very rapid, and second, a superimposed drift which extends over periods of several hours. The second effect is due to the migration of oil-soluble components across the interface, and a continuous change in the composition of the two phases. Since the two rates are quite different, Powney and Addison have been able to control the age of the surface so as to obtain results due to true adsorption which were reproducible to within one per cent. Davis and Bartell[12] have made interfacial tension measurements of sodium laurate solutions against n-heptane by the pendant drop method. They showed that the extended decrease with time of the interfacial tensions of partially hydrolyzed laurate solutions against n-heptane was due to migration of fatty acid across the interface.

2. Foaming Test

The foaming properties of the solutions were determined by the "pour foam" test described by Ross and Miles.[13] It consists of allowing 200 cc. of solution to fall through an orifice of fixed diameter into a long cylinder containing 50 cc. of solution. The height of the foam, which is taken as proportional to the volume of foam, was measured initially and at intervals up to 15 minutes. The foam produced is subjected to the destructive forces embodied in the action of the falling droplets upon the foam already formed. In most cases, the height could be reproduced within about ± 3 mm. The apparatus is provided with means for protecting the foam from evaporation and from thermal shock.

In this test, the time for foam formation is about twenty seconds. If a foam is produced in some other way, such as agitation or by bubbling gas, where the time for foam formation is much greater, different amounts of foam may be obtained at low concentrations of the detergent than in the "pour foam" test.

3. Canvas Disk Wetting Test

The sinking time of a No. 6 Mount Vernon duck canvas disk, 1 inch in diameter, was determined. 75 cc. of the solution to be tested was placed in a 150 cc. beaker, and the disk dropped in the solution so that it was in a horizontal position. The time required for the disk to sink

[12] Davis, J. K., & F. E. Bartell. J. Phys. Chem. 47: 40. 1943.
[13] [illegible], J., & G. D. Miles. Oil and Soap 18: 99. 1941.

beneath the liquid surface is designated as the wetting time. Each test was repeated several times until consistent values were obtained. The temperature of the solution was measured at the beginning and end of the test. In our tests, we did not find differences in the mechanism of wetting referred to by R. R. Ackley[13a] and consequently the values obtained are comparable for the solutions of the sodium alcohol sulfates tested. When the wetting time was greater than 300 seconds, the reproducibility became poorer so that long wetting times could not be considered significant on a quantitative basis.

4. Relative Detergency Test

The detergency tests were carried out by washing single pieces of cotton containing standard soil for 30 minutes in the Launder-Ometer.[14] The soil on the cotton was prepared by a reproducible procedure and contained Oil Dag (graphite and mineral oil) and cottonseed oil.[15] The procedure was a modification of the one suggested by Appel, Smith and Christison,[14] where the relative light reflection of the samples was determined with a Lange Universal reflectometer. The zero on the scale was set with the unsoiled cloth, and 100 units were set by a standard black surface. The units of soil removed refers to this arbitrary scale of light reflection from the cloth sample

The reproducibility of the modified Launder-Ometer test is approximately ± 15% when good detergency is obtained. For lower detergencies, the variations are larger.

This modified procedure was considered adequate to indicate significant differences in relative detergency when limited amounts of materials are available. Reproducibility of about ± 6% has been obtained on the basis of many detergency tests carried out in an electric washing machine where ten gallons of solution were used.[16]

5. Solubility Measurement

For those compounds which crystallize from aqueous solutions, solubility determinations in distilled water were made by allowing a supersaturated solution in contact with solid material to rotate for 24 hours or more in a small vial immersed in a water thermostat. This time interval was found to be more than enough to obtain equilibrium. The concentration of the saturated solution was determined by the reading

[13a] Ackley, R. R. Ann N. Y. Acad. Sci. 46 (6): 514–519. 1946.
[14] Appel, W. D., W. C. Smith & H. Christison. Am. Dyestuff Reptr 17: 679. 1928.
[15] Van Zile, B. S. Oil and Soap 20: 55. 1943.
[16] Woodhead, J. A., F. T. Vitale, & A. J. Frantz. Oil and Soap 21: 333. 1944.

obtained on a Zeiss interferometer which was calibrated with solutions of known concentrations. For those compounds which do not crystallize from water, no unambiguous criterion of solubility could be taken.

6. Measurement of pH of Soap Solutions

pH adjustments were made by the use of dilute solutions of sodium hydroxide or hydrochloric acid. Foam stabilities were frequently checked by raising and re-lowering the pH to be sure that the observed effects were attributable to pH changes and not to salt effects.

In the course of this work,[4] it was observed that the rate of recovery of the foam stability to the original value was influenced by the history of the pH changes made on the solution. If the pH was dropped and then raised, the recovery was slower than in instances where the change was from high to low pH.

Measurements of pH were made using the Beckman Model G pH meter equipped with a type E glass electrode, which is stated to be relatively free from sodium-ion error at high pH values.

IV. DISCUSSION OF DATA ON SODIUM ALCOHOL SULFATES

1. Surface Tension

The surface tension data as a function of concentration are presented in FIGURES 1–4. In most cases, the temperature of the solutions was 25° C., but, when the compound had low solubility, 40° C. was chosen.

An indication of the kind of changes obtained is given in FIGURE 5 where the surface tensions for solutions at a concentration of 4×10^{-3} molal are shown for an increasing number of carbon atoms.

The surface tension of aqueous solutions of alcohol sulfates varies in a regular way toward more pronounced lowering of surface tension as the length of the hydrocarbon chain is increased for the symmetrical series, as well as when the sulfate group is moved toward the symmetrical position for the series with 15 carbon atoms.

The ratios of the concentrations of successive members of the sodium salts of alcohol sulfates required to show the same lowering in surface tension are fairly constant for the secondary alcohol sulfates. These ratios have been interpreted as an indication of the increase in the work

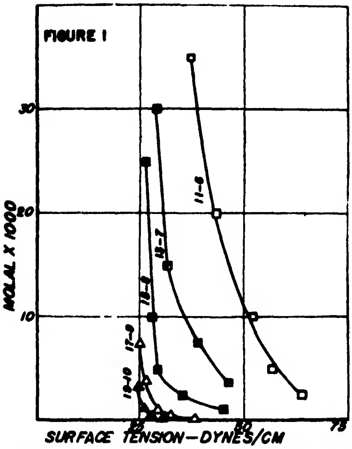


Figure 1 Surface tension vs concentration for symmetrical sodium alcohol sulfates (Dreger *et al* [1])

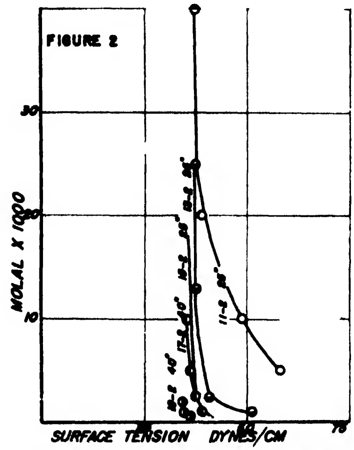


Figure 2 Surface tension vs concentration for sodium secondary alcohol sulfates (SO_4Na on second carbon atom) (Dreger *et al* [1])

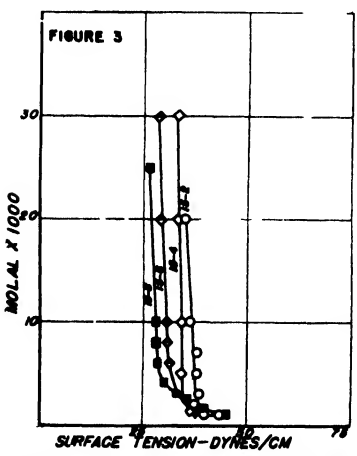


Figure 3 Surface tension vs concentration for sodium secondary pentadecanol sulfates (Dreger *et al* [1])

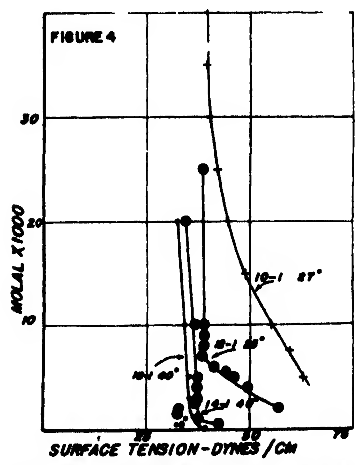


Figure 4 Surface tension vs concentration for sodium normal alkyl sulfates (Dreger *et al* [1])

done when a molecule passes from the interior to the surface layer for each additional CH_2 group Apparently, the behavior of the salts of secondary alcohol sulfates approximates Traube's rule, but this does not seem to be true for the salts of primary alcohol sulfates.

2. Interfacial Tension

The interfacial tensions against benzene are reported for solutions of sodium alcohol sulfates[17] in FIGURES 6–8. The coördinates are at right angles to those shown in FIGURES 1–4 The lowering of interfacial tension becomes greater as we increase the number of carbon atoms in a particular series. However, when the interfacial tension of solutions of the alcohol sulfates of the symmetrical series are compared with the corresponding ones of the series with the sulfate group on the second carbon atom, it is evident that the breaks in the curves come at much higher concentrations in the symmetrical series (15–8 at 3.5 × 10^{-3}M and 15–2 at 6 × 10^{-4}M). Furthermore, at higher concentrations, the

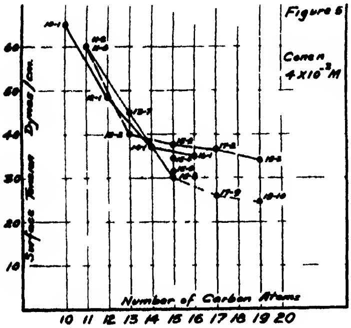


FIGURE 5 Surface tension vs number of carbon atoms for aqueous solutions of sodium alcohol sulfates (Dreger *et al* [1])

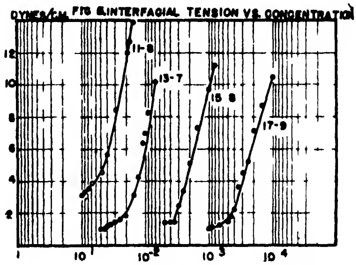


FIGURE 6 Interfacial tension (water/benzene) vs concentration for symmetrical sodium alcohol sulfates (Gerecht [17])

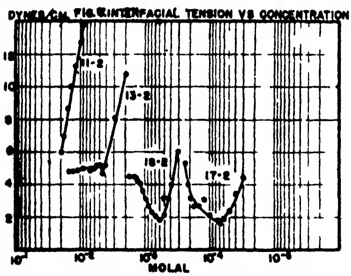


FIGURE 7 Interfacial tension (water/benzene) vs concentration for sodium secondary alcohol sulfates (SO_4Na on second carbon atom) (Gerecht [17])

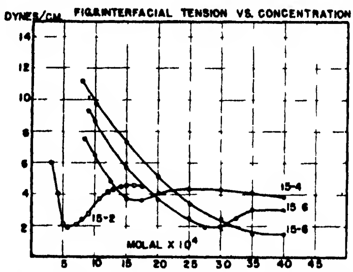


FIGURE 8 Interfacial tension (water/benzene) vs concentration for sodium secondary pentadecanol sulfates (Gerecht [17])

[17] Data obtained by J. F. Gerecht.

interfacial tensions reach lower values in the symmetrical series than in the series with the sulfate group on the second carbon atom. These differences are similar to the corresponding changes for surface tension.

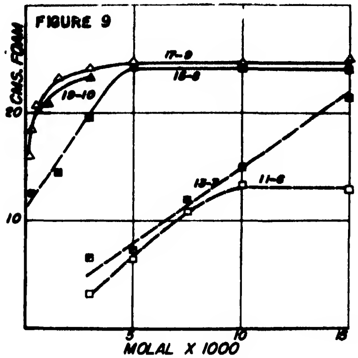


FIGURE 9 Centimeters of foam vs concentration for symmetrical sodium alcohol sulfates (Dreger *et al* [1])

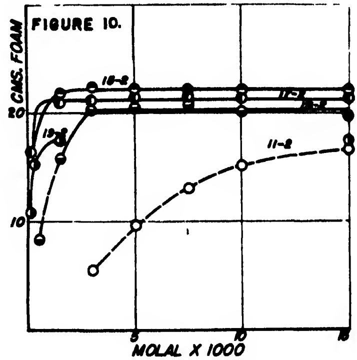


FIGURE 10 Centimeters of foam vs. concentration for sodium secondary alcohol sulfates (SO_4Na on second carbon atom) (Dreger *et al* [1])

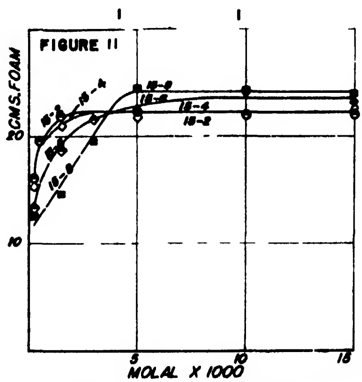


FIGURE 11. Centimeters of foam vs concentration for sodium secondary pentadecanol sulfates (Dreger *et al.*[1])

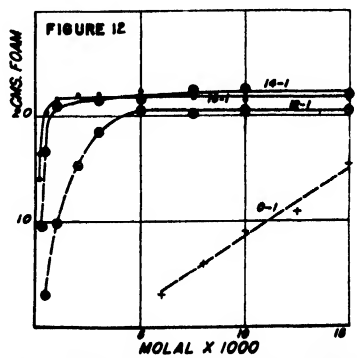


FIGURE 12. Centimeters of foam vs. concentration for sodium normal alkyl sulfates. (Dreger *et al* [1])

3. Foaming Test

FIGURES 9–12 show the initial foam height and dotted lines indicate where the foam was broken down to one-half or less of the initial height within five minutes.

For the sodium sec-alcohol sulfates with the sulfate group in the symmetrical position, the largest volume of stable foam is obtained with sodium heptadecane-9-sulfate (17–9). For solutions of sodium

sec-alcohol sulfates with the sulfate group on the second carbon atom, the foaming properties improve with increasing length of the carbon chain up to 15-carbon compound and then tend to decrease with further increase in the length of the carbon chain. For the sec-alcohol sulfates with 15 carbon atoms, the foaming properties improve as the sulfate group is shifted toward the symmetrical position. The foam of the sodium n-alcohol sulfates may be compared with that of the sodium sec-alcohol sulfates with the sulfate in the 2 position. The foaming

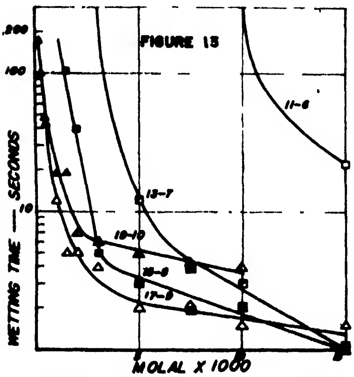


FIGURE 13. Canvas disk wetting time vs. concentration for symmetrical sodium alcohol sulfates. (Dreger *et al.*[1])

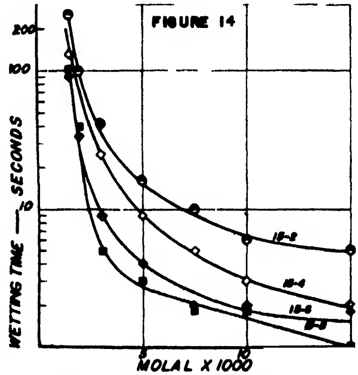


FIGURE 14. Canvas disk wetting time vs. concentration for sodium secondary pentadecanol sulfates. (Dreger *et al.*[1])

properties for these compounds increase in the following order, 10–1 < 11–2 < 12–1 = 13–2 < 14–1 = 15–2. They then decrease with further increase in the length of the carbon chain in the following order, 16–1 > 17–2 > 19–2.

4. Canvas Disk Wetting Time

Typical results of the canvas disk tests are given in FIGURES 13–14, and for a concentration of 7.5 × $^{-3}$ molal are shown in FIGURE 15. The sodium sec-alcohol sulfates showed a shorter time as the position of the sulfate group was removed farther from the end of the hydrocarbon chain. The time decreases from 15–2 to 15–4 to 15–6 to 15–8. Also, the series of alcohol sulfates 13–7, 19–10, 15–8, 17–9, for most of the concentrations tested, gave shorter times than the corresponding compounds with the sulfate group in the 2 position. It has been indicated that, for any given position of the polar group, an optimum molecular weight is associated with the fastest canvas disk test and the greatest foam stability.

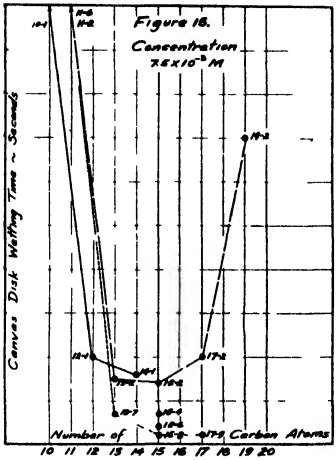


FIGURE 15. Canvas disk wetting time vs. number of carbon atoms for aqueous solutions of sodium alcohol sulfates. (Dreger *et al* [1])

5. Solubility

TABLE 1 summarizes the solubility determinations for pure sodium primary and secondary alcohol sulfates. These data[1] show marked increases in solubilities with decreasing numbers of carbon atoms and indicate sharp increases with temperature, comparable to those obtained for soaps and alkyl sulfonates.

TABLE 1

SOLUBILITY OF SODIUM SALTS OF PRIMARY AND SECONDARY ALCOHOL SULFATES[a]

	20° C.		25° C.		30° C.		35° C.		40° C.	
Compound	M × 10⁴	%	M × 10⁴	%	M × 10⁴	%	M × 10⁴	%	M × 10⁴	%
12–1	6,800	19.6	10,000	28.8	..	..	..	.	..	.
14–1		..	75	0.237	800	2.53	1,500	4.7		
16–1	..	..	.	..	9	0.03		..		
11–2	4,900	13.5		..	12,000	33.0	..	..		.
13–2	255	0.77	1,370	4.15	3,900	11 8				.
15–2	..	..	74	0.244	350	1.16	.	..	4,000	13.2
17–2	..	.	4	0.014	12	0.043			1,600	5.74
19–2				..	6	0.02		..	21	0.081

[a] $M \times 10^4 = 10^4 \times$ (molal concentration) of saturated solution; % = grams per 100 cc. of saturated solution. Temperatures are within ± 0.1° C. (From Dreger et al.[1])

6. Detergency

We may summarize the indications obtained from the detergency tests,[1] which were carried out in the modified form to show the direction of the changes obtained when limited amounts of material are available.

These tests indicate that, when other factors are the same, the nearer the polar group is to the end of a straight-chain alcohol sulfate, the better the detergency. For the series of compounds studied, an increase in molecular weight increases the detersive properties up to a certain point, and beyond this, a further increase in molecular weight serves mainly to increase the range of detergency as a function of concentration. For the compounds with 17 and 19 carbon atoms, much lower concentrations show more significant detersive properties than for the compounds with fewer carbon atoms.

Our data show that better wetting, as well as foam stability, is obtained for the compounds containing the sulfate group further away from the end of the hydrocarbon chain. Furthermore, for any given position of the polar group, an optimum molecular weight is associated with the fastest wetting and greatest foam stability. This is similar to the results reported by Wilkes and Wickert[18] for branched-chain compounds. On the other hand, the detergency tests show no such optimum within the limits set by decreasing solubilities for the higher members of the series. We believe that these differences may be due to the importance of the rates of attainment of steady states in the surface active properties for the wetting and foaming tests; whereas, for detergency tests, these particular factors may be relatively unimportant.

Our study indicates that solubilities, as well as the molecular structure and molecular weight, determine the limitations in the surface active properties of paraffin-chain salts. For the compounds studied, if the length of the hydrocarbon chain is not sufficient (less than 11 carbon atoms), the surface active properties are not very great; if the solubilities are too low, which is the case for the compounds with higher molecular weights, the effectiveness of a wetting agent, foaming agent, or detergent is also restricted. Furthermore, the low solubility of the higher members of the series prevents comparisons at the higher concentrations which are required to obtain marked surface active properties for lower members of the same homologous series.

The surface tensions of sodium decyl sulfate (10–1), sodium unde-

[18] Wilkes, B. G., & J. N. Wickert. Ind. Eng. Chem. 29: 1234. 1937.

cane-2-sulfate (11–2), sodium undecane-6-sulfate (11–6), and sodium tridecane-7-sulfate (13–7) as previously noted, show a gradual increase on dilution. This distinguishes these curves from those for the compounds where the surface tension changes abruptly at relatively low concentrations. The compounds which give the first type of curve do not show appreciable detersive properties, but the latter group with sharp breaks in the curves does show significant detersive properties. However, it must be emphasized that since numerous other factors are involved in detergency, it is by no means a general rule that either surface or interfacial tension alone will be an adequate guide to the detersive properties of a solution. Furthermore, a particular type of detergency test, such as the one described, cannot be expected to provide an absolute evaluation which will correspond to a variety of conditions.

7. Salt Effects

Salt effects were obtained which enhanced surface tension lowering, foam height, wetting test and detergency at low concentrations. These effects increase chiefly with the charge type of the cation. The values obtained are comparable but not greater than can be obtained at higher concentrations of detergent without added salt.

Aickin[19] has measured the interfacial tension against oils of solutions of sodium secondary alcohol sulfates prepared from an olefine fraction by reaction with sulfuric acid followed by neutralization. He noted that the addition of electrolytes reduces both the interfacial tension and the concentration at which the curves show a sharp break. The monovalent cations showed effects which followed a lyotropic series. Aickin notes that the ions which approach closer to the interfacial film, that is, the least hydrated ions, have the greatest influence on the interfacial tension of the system.

V. MINIMA IN SURFACE TENSION CONCENTRATION

The observation of minima in the surface tension-concentration curves for aqueous solutions of various surface-active materials has been the cause of considerable conjecture. Powney and Addison,[20] in particular, found well-defined minima in the surface tension of solutions of sodium salts of primary alcohol sulfates containing from twelve to eighteen carbon atoms. Lottermoser and Stoll,[21] show two

[19] **Aickin, R. G.** J. Soc Dyers and Colourists **60**: 36 1944.
[20] **Powney, J., & C. C. Addison.** Trans. Faraday Soc. **33**: 1243 1937.
[21] **Lottermoser, A., & F. Stoll.** Kolloid Z. **63**: 49. 1933.

minima for surface tension concentration curves of sodium cetyl sulfate solutions. Many other instances of minima in such curves have been cited. These "anomalous" results have presented apparent disagreement with Gibbs' adsorption equation. The hypotheses proposed to reconcile such incongruities have been frequently discussed.[2, 3, 22]

Miles and Shedlovsky[2] have shown that minima in certain surface tension-concentration curves can be attributed to the presence of at

FIGURE 16.

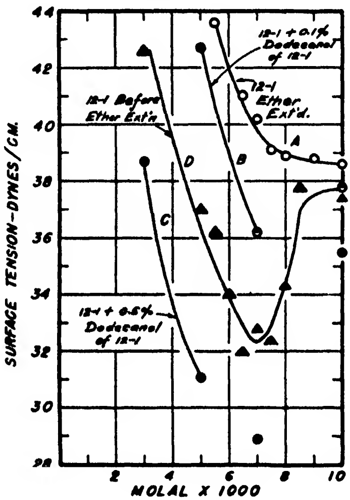


FIGURE 16. Effect of dodecanol upon the surface tension of solutions of sodium dodecyl sulfate (12–1). (Miles and Shedlovsky[2])

least two surface active substances in the same solution. A typical example is shown in FIGURE 16 where Curve A for pure sodium dodecyl sulfate, which had been extracted with ethyl ether for 36 hours in a Sohxlet apparatus, does not show any minimum. Curve D shows a marked minimum for solutions of the same material before the final extraction with ether. Curves B and C show progressively more pronounced minima as larger amounts of dodecanol are added. This shows the effect obtained when two surface-active materials are present in the solution and one of these is only slightly soluble in water. These results are explained by a greater adsorption of dodecanol than corre-

---

[22] McBain, J. W. Colloid Chemistry. 5: 102. Edited by J. Alexander. Rheinhold Publishing Co. New York, N. Y. 1944.

sponds to that for the minimum in the surface tension of the solutions shown in FIGURE 16. It was also shown that minima in surface tension-concentration curves were obtained for sodium lauryl sulfate containing a small amount of a higher molecular weight homologue. In the latter case, it was postulated that the minima were due to a salt effect of the principal constituent, sodium lauryl sulfate, on a small amount of the higher molecular weight homologue.

It is suggested that a critical examination of the material which show minima in surface tension-concentration curves, might reveal that their "anomalous" behavior was a result of various impurities or hydrolysis products rather than the characteristic of a single surface active component.

Minima in surface or interfacial tension-concentration curves have been produced by deliberate "contamination" of solutions of pure, surface active, primary or secondary alcohol sulfates with a second surface-active material.

Selective adsorption experiments at air liquid interfaces[3] were carried out with foam of uniform bubble size. In each case, it was assumed that the interfacial area was proportional to the height of the foam column. The interfacial area was estimated to be approximately 2×10^4 cm.^{2} and the same for all the foam adsorption experiments. The foams produced were sufficiently stable to prevent any appreciable coalescence of the bubbles during the experiment.

FIGURE 17, Curve 1, represents the surface tension as a function of concentration for pure sodium lauryl sulfate which had been extracted with ether for 36 hours in a Soxhlet apparatus. Curve 2 illustrates the minimum obtained when sodium lauryl sulfate is contaminated with 0.5 per cent lauryl alcohol, on the total solids basis.

A 0.015 molal solution with surface tension (A) was agitated gently for several minutes with the column of foam. When this solution was diluted to 0.0075 molal, the surface tension was (A^1). A 0.010 molal solution with surface tension (B), after following the same foam adsorption procedure and diluting to 0.0075 molal, gave a surface tension (B^1). (C^1) was obtained by the same steps as (A^1) and (B^1), except that the initial and final concentrations were the same (0.0075 molal). The same foam adsorption procedure was followed for a 0.005 molal solution with an initial surface tension (D) which was increased to (D^1). When the concentration of this solution was increased by the addition of pure sodium lauryl sulfate, a surface tension (D^{11}) was obtained. The above typical adsorption experiments indicate that in the surface tension-concentration curves considered, the minima occur

at bulk concentrations where the relative surface concentration of the minor component is at a maximum. Selective adsorption experiments at benzene-liquid interface have led to the same conclusion.

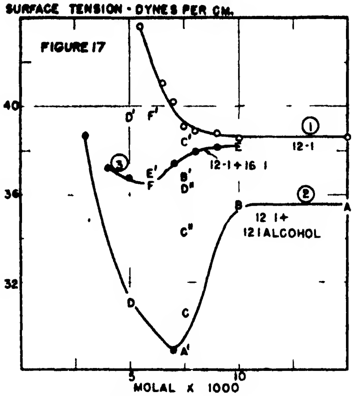


FIGURE 17 Effect of foam extraction on surface tension of sodium dodecyl sulfate (12-1) + 0.5% dodecyl alcohol and sodium dodecyl sulfate (12-1) + 1% sodium hexadecyl sulfate (16 1) (Miles[3])

VI. PROPERTIES OF SOAP SOLUTIONS

1. Surface and Interfacial Tension

The complex nature of the surface tension and interfacial tension data for soap solutions is given by Powney and Addison.[5, 6] They point out that, for laurates, very low minimum values of surface tension could be obtained only if precautions were taken to eliminate all traces of alkali other than that produced by the pure soap. For example, 0.1 per cent potassium laurate solution has a surface tension of 23 dynes per cm. (pH 7.8). If the pH of this solution is raised to 9.5, the surface tension rises to 53 dynes per cm. (FIGURE 18).

Many previous studies did not give concordant results, due largely to effects of hydrolysis. Small amounts of hydrolysis in soap solutions are considerably magnified by the large difference in adsorption of soap anion, acid soap, and fatty acid. Powney emphasizes the magnitude of the effect of carbon dioxide on the surface tension of soap solutions. He points out that a 0.1 per cent sodium laurate solution decreased from pH = 9.1 to pH = 7.7 on exposure to air, the solution eventually becoming turbid. Similar effects may be expected for solutions of

potassium laurate and, from FIGURE 18, we can see that this would cause a very considerable drop in surface tension after exposure to the atmosphere.

The changes in surface and interfacial tension of solutions of sodium laurate, myristate, palmitate and stearate give an indication of the effect of increasing molecular weight in a homologous series (FIGURES 19–22).

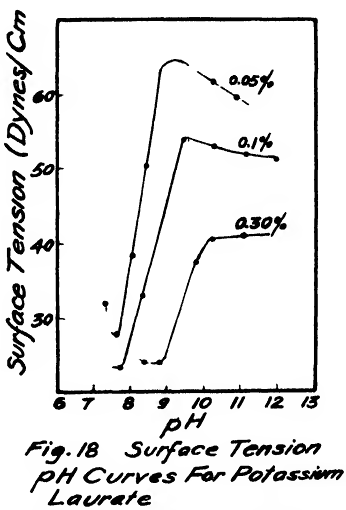


FIGURE 18 Surface tension—pH curves for potassium laurate solutions (Powney[5])

The surface tension curves under conditions of suppressed hydrolysis presented in FIGURE 20 shows a progressive lowering in the values with increasing molecular weight.

The curves for interfacial tension against xylene of sodium laurate, myristate, palmitate and stearate show a sharp break in the curve which becomes less pronounced with increasing chain length. For sodium stearate, the curve (FIGURE 21) tends to lie between the sodium laurate and sodium myristate curves. The differences are explained on the basis of increasing degree of hydrolysis of the soap solutions with increasing chain length.

In interfacial tension measurements, it has been considered[6] that the surface activity at an oil-solution interface is due to the single long-

chain ions, and at an air-solution interface the acid soap is more surface active than are the single long chain ions. Therefore, at high

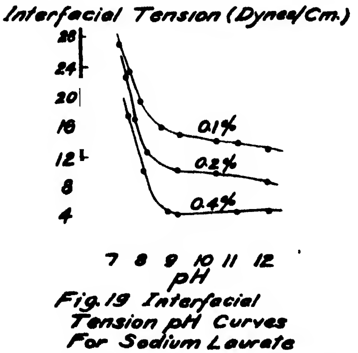


FIGURE 19. Interfacial tension (water/xylene)-pH curves for sodium laurate. (Powney and Addison.[6])

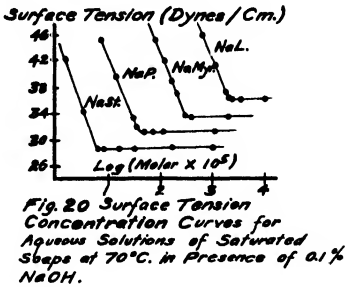


FIGURE 20. Surface tension-concentration curves for aqueous solutions of saturated soaps at 70° c. in presence of 0.10% NaOH. (Powney and Addison.[6])

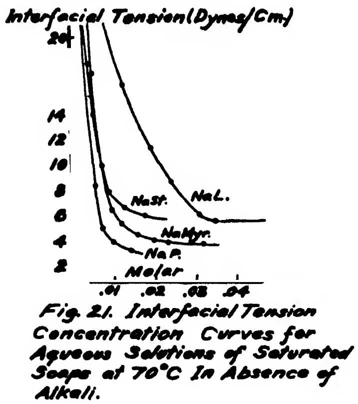


FIGURE 21. Interfacial tension (water/xylene)-concentration curves for aqueous solutions of saturated soaps at 70° C. in absence of alkali. (Powney and Addison.[6])

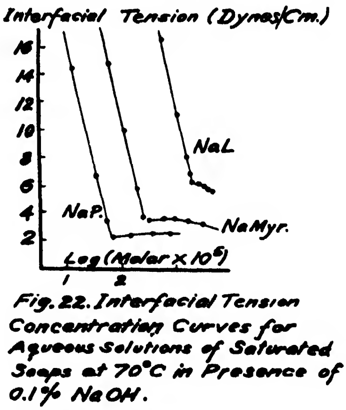


FIGURE 22. Interfacial tension (water/xylene)-concentration curves for aqueous solutions of saturated soaps at 70° C. in presence of 0.1% NaOH. (Powney and Addison.[6])

pH values the interfacial tension concentration curves are displaced towards lower concentrations with only a slight alteration in the maximum lowering of the interfacial tension, while increase in the pH of the solution tends to increase the surface tension.

2. Foaming Test

Miles and J. Ross[4] have shown the effect of concentration and changes in pH on the relative foam stability of solutions of pure sodium soaps (FIGURES 23–25).

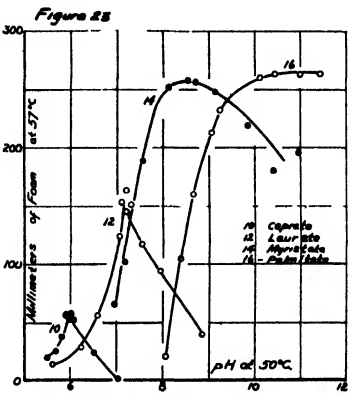


FIGURE 23 Foam stability of 0.1 per cent solutions of pure sodium salts of saturated fatty acids as a function of pH at 57° C (Miles and Ross[4])

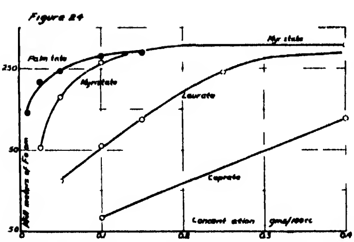


FIGURE 24 Foam stability of solutions of pure sodium salts of saturated fatty acids at 57° C as a function of concentration where each solution was adjusted to the pH for maximum foam stability (Miles and Ross[4])

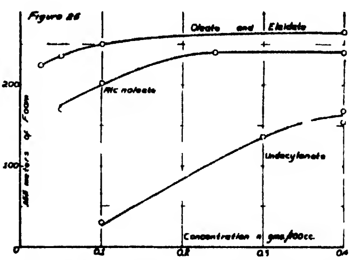


FIGURE 25 Foam stability of solutions of pure sodium salts of unsaturated fatty acids at 57° C as a function of concentration where each solution was adjusted to the pH for maximum foam stability (Miles and Ross[4])

One of the fundamental distinctions between the properties of soaps and those of other detergents, such as the sodium alcohol sulfates, is that soap solutions hydrolyze, while synthetic detergents which are salts of strong acids do not hydrolyze appreciably at room temperatures This is a complicating feature of any study of the surface-active properties of soap solutions because at least two of the hydrolysis products

in such solutions are surface-active; namely, the undissociated fatty acid and the fatty acid anion. There is a difference in apparent weakness of the various fatty acids, which is probably a manifestation of differences in their solubility, since the pK values have been reported[23] to be between 4.9 and 5.0 for the saturated fatty acids. This is illustrated in FIGURE 23, where the pH determines the balance between the relative amounts of the various surface-active components. This is reflected in the changes in stability of the foam as a function of changes in the pH of the solutions. If the pH of the sodium caprate or sodium laurate solutions was raised to a point where the concentration of the undissociated fatty acid approached zero, the foam stability was also nearly zero. Powney's[5] analogous surface tension behavior has been noted for potassium laurate solutions when the pH was increased.

Cupples[24] found related changes in contact angle, interfacial tension and surface tension for soaps as a function of the pH.

The relative foam stability of the soap solutions was compared as a function of concentration where each solution was adjusted to the pH for maximum foam stability. These data are shown in FIGURES 24 and 25. The characteristic increase in foam stability associated with increasing molecular weight in a homologous series is indicated in these curves.

Mixtures of the sodium laurate and palmitate have been shown to give a greater foam stability than either solution alone. For example, at pH 8.5, 0.1 per cent sodium laurate solution gave a foam height of 55 mms. and 0.01 per cent sodium palmitate, a foam height of 25 mms. On the other hand, a mixture of these two solutions gave a foam height of 185 mms.

An interesting feature in the behavior of these materials is the important role played by the fatty acids in stabilizing the foams of certain solutions. It seems probable that the films are mixtures of acid anion and fatty acid This suggests that other combinations of two surface active materials, neither of which produces stable foams, when blended, might yield a relatively more stable foam. The number of surface active materials which do not produce stable foams is fairly large, and therefore a further clue as to the desirable nature of each member of the pair was sought in the prototype mixture of sodium laurate-lauric acid. The laurate anion might be considered as a water-soluble surface-active ion in which the hydrophilic character is so dominant that adsorption is inadequate for production of stable foam.

[23] Jukes, T. H., & C. L. A. Schmidt. J. Biol. Chem 110: 9 1935
[24] Cupples, H. L. Ind. Eng. Chem. 29: 924 1937.

The lauric acid, on the other hand, is a member of a series of homologues in which none exhibit foam-stabilizing properties and those members which are appreciably surface active are only slightly water soluble. It was found that lauryl alcohol would stabilize the foam of 0.1 per cent sodium laurate solutions at pH 10 and lauric, myristic, or palmitic acid would stabilize the otherwise unstable foam of 0.2 per cent sodium decyl sulfate solutions at pH 4. In each case, the solutions were saturated with respect to the least soluble component. With each of the acids, a definite temperature had to be exceeded before any foam could be produced. Washing the stabilized foams with solutions of either one of the components alone invariably led to foam collapse.

VII. SUMMARY

The preparation and properties involving surface activity are described for alternate members of a homologous and an isomeric series of purified sodium salts of secondary alcohol sulfates and of sodium salts of primary alcohol sulfates with 10, 12, 14 and 16 carbon atoms The surface tension, interfacial tension (benzene/water), foaming, wetting and detersive properties of solutions of these compounds are reported. The solubilities of those compounds which can be easily crystallized are measured at 5° intervals from 20° to 40° C. The data are discussed from the point of view of correlating changes in the properties of the compounds when their structures and molecular weights are changed.

The importance of considering the effects of small amounts of impurities which lead to minima in surface tension-concentration curves is discussed. It is shown that such minima can be attributed to the presence of at least two surface active substances in the same solution. In all cases, attempts at selective adsorption at either an air-liquid or benzene-liquid interface have indicated these minima to occur at bulk concentrations where the relative surface concentration of the minor component is at a maximum.

The changes in surface and interfacial tension of soap solutions as a function of pH are discussed. At an air/solution interface the acid soaps are more surface active than the long-chain ions, whereas at the xylene/solution interface, the surface activity is due to the single long-chain ions. In accord with this concept, when the pH of soap solutions is increased, the surface tension increases, but the interfacial tension (xylene/solution) decreases. The effects of changes in concentrations and in pH upon the relative foam stabilities are re-

ported for solutions of sodium soaps. The surface and interfacial tension data can be useful in explaining these foam stabilities. Mixtures of sodium laurate and sodium palmitate give greater foam stability than either one alone. Other examples of mixtures which give better foams than either constituent alone are lauryl alcohol and sodium laurate solution at pH 10 and lauric, myristic or palmitic acid added to sodium decyl sulfate solution.

SURFACE ACTIVE AGENTS IN BIOLOGY AND MEDICINE

By E. I. Valko

Onyx Oil & Chemical Company, Jersey City, New Jersey

INTRODUCTION

The bile acids, the lecithins and the glucosides, particularly the saponins, are surface active agents occurring in the living organism. The chemical structure and biological function of these compounds have been from an early date of interest to the biochemists (cf. H. Sobotka[1]). Some of these compounds have been used therapeutically by primitive people and are still in use. Sodium chaulmoograte was introduced some time ago for the treatment of leprosy, for which chaulmoogra oil has been used for more than a century in the Orient. It is interesting to note that one of the first commercially successful synthetic detergents was N-oleyl, N-methyl tauride (A) possessing a chemical structure closely related to that of the natural taurocholic acid (B).

$$\underset{C_{17}H_{33}CON(CH_3)CH_2CH_2SO_3^-}{\text{(A)}} \qquad \underset{C_{23}H_{39}O_3CONHCH_2CH_2SO_3^-}{\text{(B)}}$$

Synthetic surface active agents for biochemical studies and medicinal purposes were used rather infrequently until recently, yet such agents have been prepared by biochemists by methods which only later attained importance for commercial production. As far back as 1911, Hunt and Taveau,[2] of the United States Public Health Service, prepared 79 choline derivatives in order to investigate their effect on the blood pressure. Among these was the condensation product of palmitic acid with choline chloride.[3, 4, 5, 6]

$$C_{15}H_{33}COOCH_2CH_2N(CH)_3^+Cl^-$$

Twenty years later, the same principle, namely esterification of fatty acids with an alcohol containing an ionic group in the molecule, was successfully utilized for the commercial production of surface active agents.

[1] Sobotka, H. Ann. N. Y. Acad. Sci. 46(6): 508, 509. 1946.
[2] Hunt, R., & R. de M. Taveau. Hygienic Laboratory, Bulletin 12. 1911.
[3] Fourneau, E., & H. J. Page. Bull. Soc. Chim. (IV) 15: 544. 1914.
[4] Bergmann, M., & E. Sabetay. Z. Physiol. Chem. 137: 47. 1924.
[5] Karrer, P. Helv. Chim. Acta. 5: 469. 1922.
[6] Hartmann, M., & H. Kaegi. Z. angew. Chem. 41: 127. 1928.

Synthetic surface active compounds have been substituted for soaps as dispersing and emulsifying agents in medicinal preparations. While they seem to offer definite advantages for some purposes, they can not generally be substituted for soaps until the required studies showing lack of cumulative irritation and toxicity have been completed (Calvery[7]).

Some data on the toxicological properties of a few surface agents are available.[8, 9, 10, 11, 12]

It should be briefly mentioned that the conversion of therapeutic agents to surface active compounds by the introduction of paraffin chains into the molecule, e.g., into sulfanilamide[13, 14, 15] or quinine, arsenicals, acridine[16, 17] did not as yet yield practical results, even though experiments *in vitro* have been promising.

Mainly due to the discovery of the denaturing effect of surface active agents on proteins by Anson[18] and of the extremely high germicidal power exhibited by a certain class of these compounds (Hartman & Kaegi,[6] Domagk[19]), biochemists, in recent years, are showing an increasing interest in this field of chemistry.

COMBINATION OF SURFACE ACTIVE IONS WITH PROTEINS

The fact is definitely established that surface active ions combine with or are adsorbed by proteins. Many of the biological effects of surface active agents may be ascribed to this combination. It is surprising that, until recently, no investigation has been made with the sole purpose of measuring the combination of surface active agents with proteins. Most of the data available were obtained from investigations carried out in the study of the role of surface active agents in dyeing and scouring of wool. Although augmented by recent investigations, the experimental information is still meager.

[7] **Calvery, H. O.** 107th Meeting Am. Chemical Soc. Cleveland. 1944.
[8] **Epstein, S., A. H. Throndson, W. M. Dock & M. L. Tainter.** J. Am. Dental Assoc. **26:** 1461. 1939.
[9] **Hatton, E. H., L. S. Fosdick & J. J. Calandra.** Dental Research **19:** 87. 1940
[10] **Smyth, H. F., Jr., J. Seaton & L. J. Fischer.** Indus. Hyg. & Toxicol. **23:** 478. 1941.
[11] **Fogelson, S. J., & D. E. Shoch.** Archives Int. Med. **73:** 212. 1944.
[12] **Kirsner, J. B., & E. A. Wolff.** Gastroenterology **2:** 93. 1944.
[13] **Behnisch, R., F. Mietzsch & J. Klarer.** British Patent No. **474,423.**
[14] **Crossley, M. L., E. H. Northey & M. E. Hultquist.** J. Am. Chem. Soc. **61:** 21950. 1939.
[15] **Arnold, H., E. Kelmert, Th. Möbus, E. Prigge, H. Rauen & Th. Wagner-Jau-**

J. Chem. Soc. **1:** 1. 1939.
Bergmann. J. Chem. Soc. **1:** 576. 1940.
1939.
Domagk, G. Deut. med. Wochschr. **61:** 829. 1935.

A. Theoretical

Combination of ions with proteins involves two kinds of molecular forces: those of specific attraction or intrinsic affinity, and those of electrostatic coulombic forces arising from the fact that the protein molecules as well as the adsorbed ions carry free electric charges

We shall first consider the electric forces. If an isoelectric protein combines with hydrogen ions, it acquires a positive electric charge. The surface of such a protein molecule exhibits against the bulk of the solutions a positive electric potential and, therefore, the protein exerts a repulsive force against cations, especially also against the free hydrogen ions present in the solution. At the same time, it exerts an attraction force for the anions present. On the other hand, adsorption of anions will diminish the positive charge of the protein and, therefore, counteract the effect of the adsorption of hydrogen ions on the electric potential. Thus, the eletrical forces between protein and ions depend on the amount and nature of all ions present in the system

A further complication arises when the system is not homogeneous but consists of microscopically heterogeneous phases, as is the case with the adsorption of ions by insoluble proteins, for instance, those of fibers or bacteria. The adsorbed ions such as hydrogen ions, dye ions, or surface active ions do not accumulate on the microscopic surfaces, but penetrate the fibers or bacteria and are distributed through the whole cross section. These ions can not enter from the solution into the second phase without being accompanied by the equivalent amount of oppositely charged ions, called counter-ions. The accumulation of counter-ions in the second phase is opposed by their kinetic energy which tends to distribute them uniformly in the whole space available. Expressed in terms of statistical mechanics, the adsorption of the counter-ions is opposed by the law of probability. However, the accumulation of the counter-ions is promoted by the electric potential of the protein phase, which is acquired as soon as an immeasurably small excess of surface active ions is adsorbed. The final adsorption equilibrium is reached when all the forces involved, molecular attraction, electric and osmotic forces are balanced.

A quantitative expression can be formulated for the dependence of the combination of an insoluble protein with surface active ions on the concentration of these ions and their counter-ions in a simplified case. Consideration of this formula helps in understanding the more complicated cases.

A salt consisting of a surface active anion and its counter-ion is added to an insoluble protein at the isoelectric point. Let us assume,

first, that the amount of anions adsorbed is relatively small. In this case, the number of available sites in the protein is not to be regarded as a variable, but as constant. Boltzmann's law for the distribution of a compound between two kinds of sites can be applied. $c_{s,pr}$ is the

(1) $$c_{s,pr} = c_{s,sol}\, e^{\frac{-w_s}{RT}}$$

molar concentration of the surface active anion in the protein; $c_{s,sol}$ the concentration of the surface active ion in the surrounding aqueous medium; w_s is the work which is gained by transferring one mole of the surface active ion from the solution in to the protein. This work can be separated into non-coulombic energy which represents the intrinsic affinity of the surface active ion of the protein, and into the electric energy which is equal to the product of charge and electric potential difference between protein and solution, that is, ψ Fz where F is the Faraday, ψ is the electric potential and z the number of electric charges of the counter-ion. Denoting the non-coulombic energy with $w_{s,a}$, for a monovalent surface active anion we can write

(2) $$c_{s,pr} = c_{s,sol}\, e^{\frac{-w_{s,a} + F\psi}{RT}}$$

For a monovalent counter-ion (i.e. sodium) the analogous equation is valid:

(3) $$c_{Na,pr} = c_{Na,sol}\, e^{\frac{-F\psi}{RT}}$$

If we assume that no other ions but the surface active ion and the sodium ion enter the protein, the concentration of these ions in the protein must be equal to each other. It follows:

(4) $$c_{s,pr} = \sqrt{c_{Na,sol} \times c_{s,sol}}\; e^{\frac{-w_{s,a} - w}{2RT}}$$

and since the intrinsic affinity of the sodium ions for the proteins can be ignored:

(5) $$c_{s,pr} = \sqrt{c_{Na,sol} \times c_{s,sol}}\; e^{\frac{}{2RT}}$$

The electric potential does not appear in this equation[20] although it is the only reason for the dependence of the adsorbed anions on the concentration of the cations. In spite of the fact that the cations are assumed to have no specific attraction for the protein, their concentration has the same effect on the adsorption of the anions as the concentration of the anions themselves.

[20] Gilbert, G. A., & E. K. Rideal. Proc. Roy. Soc. London A 182: 335. 1944.

Equation (5) holds also for the presence of sodium chloride in the solution as long as the concentration of the chloride ions in the protein remains comparatively low. Since chloride ions have no specific attraction to the proteins and are repulsed by the electric forces, the chloride ion concentration in the solution must reach a fairly high value before a sufficient amount of chloride ions enters the protein to make the equation invalid.

According to Equation (5), addition of a salt possessing the same counter-ion as the counter-ion in the surface active compound will facilitate the adsorption of the surface active ions. This is to be expected, since, if the concentration of the counter-ion in the solution is higher, the required relative increase of their concentration in the fiber will be less.

An equation analogous to Equation (5) holds for the adsorption of a monovalent acid by the protein, if the anion has practically no intrinsic affinity for the protein. When both cation and anion possess affinity for the protein, for instance, in the case of an anionic surface active compound applied in the form of a free acid, the sum of the intrinsic affinities of both the ions will promote the adsorption as expressed by Equation (4).

A modification of the equations is required if the number of available sites in the protein can not be regarded as constant. In this case, the concentration of adsorbed ions $c_{s,\,pr}$ in the equations has to be replaced, according to the law of mass action, by the expression $\frac{\alpha}{1-\alpha}$ in which α means the fraction of occupied sites.

The simple Equation (2) explains the main characteristics of the concentration relationship involved in the interaction of insoluble, but for ions, permeable, proteins with surface active ions as well as with dye ions.[21, 20, 22] Substituting membrane potential for fiber potential which is designated by ψ, this equation becomes identical with Donnan's equation of the membrane equilibrium which has been applied to the reaction of fibrous proteins with acids and to the related phenomena of dyeing by Speakman[23] and by Elöd and Silva.[24]

B. Experimental

Meyer and Fikentscher[25] measured the adsorption of dibutyl naphthalene sulfonic acid by wool and characterized it as the formation of

[21] **Hartley, G. S., & J. W. Roe.** Trans. Faraday Soc. **36:** 101. 1940.
[22] **Gilbert, G. A.** Proc. Roy. Soc. London A **183:** 167. 1944.
[23] **Speakman, J. B.** J. Soc. Dyers Colourists **41:** 172. 1925.
[24] **Elöd, E., & E. E. Silva.** Z. phys. Chem. A. **137:** 142. 1927.
[25] **Meyer, K. H., & H. Fikentscher.** Melliand Textilber. **7:** 605. 1926.

a salt with the protein which functions here as a base. In the language of the above outlined theory, the hydrogen ions are combined with wool because their intrinsic affinity for the free amino groups of the wool (or, rather, according to the zwitter-ion theory, to the ionized carboxylic groups). The anions are adsorbed in order to balance the electric charge of the adsorbed hydrogen ions. Neville and Jeanson[26] determined the adsorption of dodecyl sulfate and N-oleyl-N-methyl tauride by wool as a function of the pH (FIGURE 1). Here again, in acidic solutions, the hydrogen ions are combined with the wool, due to their intrinsic affinity, and the anions are adsorbed because of the necessity

FIGURE 1 Adsorption of surface active anions by wool Compound A sodium lauryl sulfate Compound B sodium N-oleyl N-methyl tauride Neville and Jeanson J Phys Chem 37: 1000 1933

to balance the electric charge. However, the pH being adjusted with hydrochloric acid, the two kinds of anions, the chloride ions and the surface active ions, compete with each other, and the preferential combination of the protein with the surface active ions is due to the higher intrinsic affinity of the latter. Actually, the chloride ions, due to their higher mobility, penetrate the fibers faster and are adsorbed first, to be replaced thereafter by the slower moving surface active ions Accordingly, this combination of the surface active ions with the proteins can be regarded as a simple ionic exchange not involving electrical forces. It is significant that, even on the alkaline side of the isoelectric point, some surface active ions are adsorbed. Here, the combination occurs in opposition to the electrical forces because of the specific affinity of the surface active ions.

[26] Neville, H. A., & Ch. A. Jeanson. J. Phys Chem 37: 1000 1933

Meyer and Fikentscher,[25] Elöd and Boehme,[27] Speakman and Stott[28] and Ender and Miller[29] observed deviations between the titration curves of wool with different acids. Steinhardt, Fugitt and Harris[30] were the first to indicate the full significance and extent of these differences. FIGURE 2 represents their data obtained at 50° C., on the combination of wool protein with ten different strong acids. While the pH of half the maximum combination with hydrochloric acid is 2.3, with dodecyl

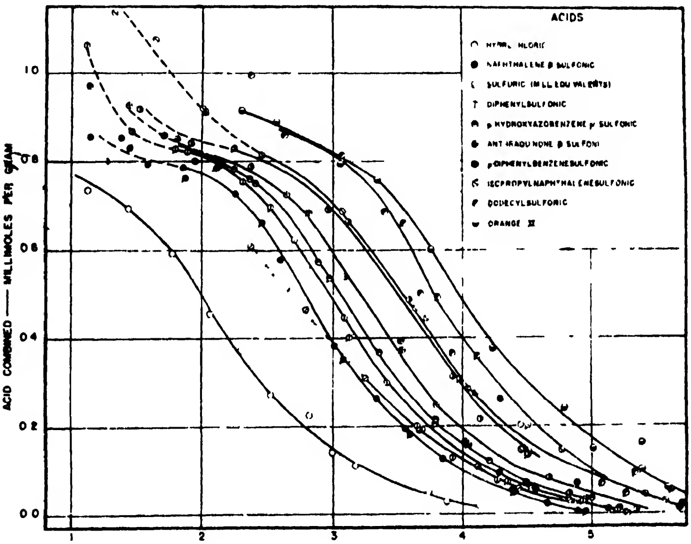


FIGURE 2. Combination of wool with 10 different strong acids as a function of pH at 50° C. Steinhardt, Fugitt & Harris. J. Research Natl. Bur. Standards **28**: 201 1942

sulfonic acid it is 3.96. Steinhardt, Fugitt and Harris assume that the protein forms undissociated salt-like compounds with the anions. In their system of dissociation equilibria, the reciprocal value of the dissociation constant of the combination of the protein with the anion is a measure of the affinity of the anion for the protein. As Gilbert and Rideal[20] showed, the quantitative treatment of Steinhardt, Fugitt, and Harris is open to criticism, since they ignored the electrostatic compulsion of the counter-ions to enter the fibers and replaced it by the im-

[27] **Elöd, E., & F. Boehme.** Melliand Textilber. **13:** 365 1932.
[28] **Speakman, J. B., & E. Stott.** Trans. Faraday Soc. **31:** 1425. 1935
[29] **Ender, W., & A. Müller.** Melliand Textilber. **18:** 663, 732, 809, 906, 991. 1937
[30] **Steinhardt, J., C. H. Fugitt, & M. J. Harris.** Research Natl. Bur. Standards **26:** 219, 293. 1941; **28:** 201. 1942.

plicit assumption of adsorption of undissociated acids. Nevertheless, their experimental material represents a most extensive comparison of the affinities of anions to wool protein and the relative order of affinities derived by them is doubtlessly correct. This order follows closely the order of molecular weight (FIGURE 3). The highest affinities are reached among the investigated acids by dodecyl sulfuric acid, dodecyl sulfonic acid and Orange II acid. Surface activity by itself seems to be unimportant for the affinity, since the surface tension of Orange II

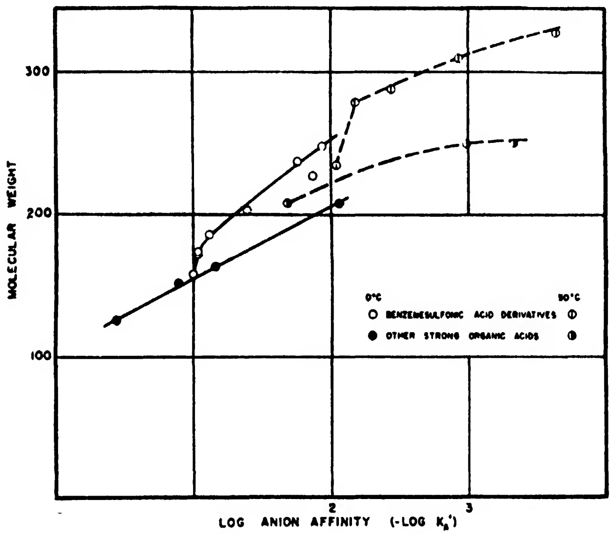


FIGURE 3. Molecular weight and values of log. affinity for anions of strong monobasic organic acids. Steinhardt, Fugitt & Harris. J. Research Natl. Bur. Standards 28: 201. 1942.

solutions is much higher than that of dodecyl sulfuric or sulfonic acid. The observations of Steinhardt, Fugitt and Harris can be readily interpreted by the above outlined theoretical consideration. Equation (4) can be applied to them by substituting H for Na and it shows that the high intrinsic affinity of the anions shifts the equilibrium in the direction of increased combination of the protein with the hydrogen ion.

It was shown by Steinhardt, Fugitt, and Harris that the titration curve of a soluble protein, egg albumin, exhibits a dependence upon the nature of the anion similar to the dependence exhibited by wool. This seems to contradict the conception that the electric fiber potential is responsible for the deviations stated above. However, using ψ

in Equation (2) for designation of the potential near a soluble protein ion or, more exactly, at a distance of its closest approach to another ion, the equation becomes identical with the fundamental equation of the Debye-Huckel theory which was also applied to the titration of soluble proteins.[31, 32] The take-up of hydrogen by soluble protein cations, similar to that by insoluble proteins, is bound to increase when their net positive charge is diminished by association with counter-ions of high affinity.

The effect of the anions of high affinity on the equilibrium between protein and hydrogen ion can also be described as a replacement of the hydroxyl ions of the proteins by these anions. The results of Steinhardt, Fugitt and Harris, mentioned above, were restricted mainly to the effect of the surface active anions on the acidic side of the isoelectric point. Previously, Stearn[33] observed that acid dyes liberate hydroxyl ions and basic dyes liberate hydrogen ions when reacted with gelatin, on both sides of the isoelectric point. Putnam and Neurath[34] recently observed a very pronounced shift of the hydrogen ion concentration due to the addition of surface active anions to an isoelectric protein. A solution of sodium dodecyl sulfate adjusted to pH 4.85 was added to serum albumen adjusted to the same pH. With an increasing amount of the detergent, the pH shifted gradually to higher values and by addition of relatively larger amounts it reached pH 6.4 (FIGURE 4). The probable explanation of this phenomenon is that, by adsorption of the dodecyl sulfate ions, the protein acquired a negative charge and thus the hydrogen ions became electrostatically attracted to it.

Only few quantitative data are available on the combination of soluble proteins with surface active ions. The combination of soluble proteins with hydrogen and chloride ions is comparatively easily estimated by electrometric measurements[35] but no simple methods are available with more complex ions. Even the determination of membrane equilibrium by using semi-permeable membranes, a method which yielded valuable information on the combination of proteins, for instance, with thiocyanate ions,[36] would meet difficulties when applied to surface active ions. These difficulties are due to the formation of ionic micelles by the surface active ions. However, Lundgren, Elam, and

[31] **Linderstrom-Lang, K.** C. R. Carlsberg Lab. **15**(7). 1924.
[32] **Cannan, R. K.** Chem. Rev. **30**: 395. 1942.
[33] **Stearn, E. A.** J. Phys. Chem. **32**: 973. 1930
[34] **Putnam, W., & H. J. Neurath.** J. Am. Chem. Soc. **66**: 692, 1992. 1944.
[35] **Pauli, Wo., & E. I. Valko.** Kolloidchemie der Eiweisskörper. T. Steinkopff. Dresden and Leipzig. 1933.
[36] **Rona, P., & H. E. Weber.** Biochem. Z. **203**: 429. 1928.

O'Connell,[37] as well as Putnam and Neurath,[38] studied the interaction of proteins with surface active anions by means of viscosity and electrophoretic measurements.

According to Lundgren, Elam and O'Connell, the electrophoretic pattern shows that when dodecyl benzyl sulfonate is added to an excess of egg albumin, a complex forms that possesses a constant ratio of protein to detergent, and the free protein forms a slower moving separate boundary. When excess detergent is added to the compound, two electrophoretic boundaries appear, and, in this case, the

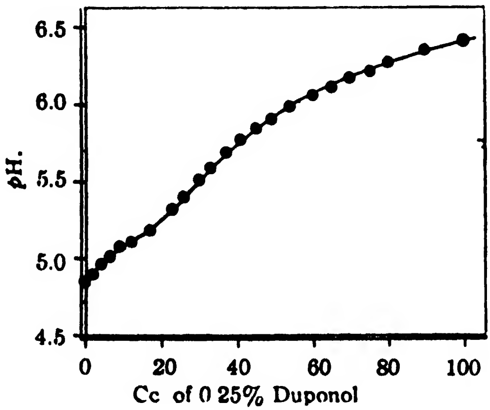


FIGURE 4 Change in pH of a salt-free 1% serum albumin solution (isoelectric point, pH 4.85) upon the addition of increments of Duponol (sodium lauryl sulfate) adjusted to pH 4 85 Original volume of serum albumin solution 100 cc. Putnam & Neurath J Am Chem Soc 66: 692 1944

slower boundary represents a complex, and the other, the free detergent. The composition of the complex is approximately constant with three parts, by weight, of protein to one part of detergent. This composition was confirmed by chemical analysis after precipitation by electrolytes.

In the mixture of sodium dodecyl sulfate and serum albumin, Putnam and Neurath could identify three electrophoretic components: (1) consisting of free protein; (2) a complex having 0.22 grams of sodium dodecyl sulfate per gram of protein; and (3) another complex having 0.42–0.45 grams of detergent per gram of protein. The composition of the latter approximately corresponds to the total acid binding capacity of the protein. It is remarkable that these findings suggest that, between pH 4.5 and pH 6.8, combination of anionic detergent and pro-

[37] Lundgren, H. P., W. Elam & R. A. O'Connell. J. Biol. Chem. 149: 183 1943
[38] Putnam, W., & H. J. Neurath. J. Am. Chem. Soc. 66: 1992. 1944.

tein is independent of pH and involves protein groups, presumably cationic, whose state of ionization does not change within that range The formation of complex (3) with a larger amount of surface active ion has been tentatively ascribed to the partial unfolding of the protein with liberation of cationic groups which, hitherto, were accessible to proteins but not to large anions.

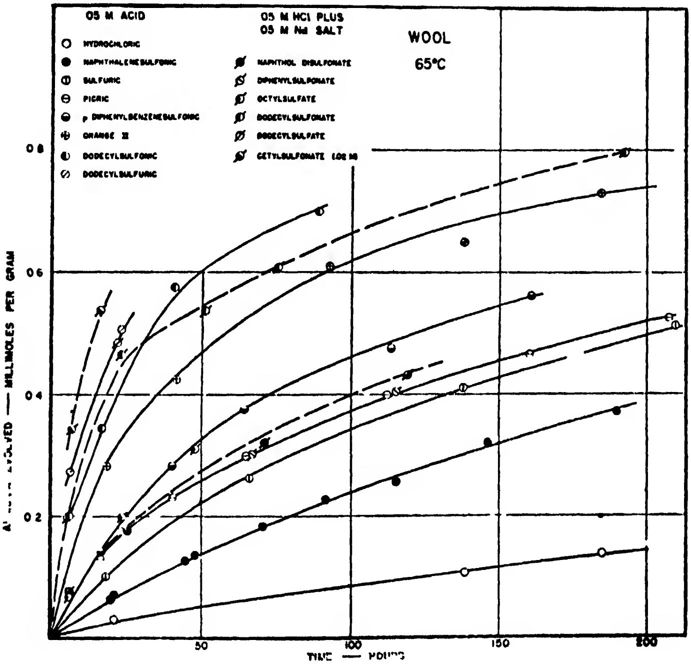


FIGURE 5 Relative effectiveness of various strong acids in hydrolyzing the primary amide bonds of wool Steinhardt & Fugitt J Research Natl Bur Standards **29**: 315 1942

Steinhardt[39] observed that the rate of hydrolysis of proteins in dilute acidic solutions increases when a small amount of anions of high affinity, for instance, dodecyl sulfate, dodecyl sulfonate, tetradecyl sulfate or Orange II is added to the solution (FIGURE 5).[40] The increase can be a hundredfold. This interesting phenomenon can be explained by the reduction of the positive electrostatic potential of the protein which has the function of a barrier against the hydrogen ions of the

[39] **Steinhardt, J.** J. Biol Chem **141**: 995. 1941.
[40] **Steinhardt, J., & C. H. Fugitt.** J. Research Natl. Bur. Standards **29**: 315. 1942.

solution. If the potential is reduced by 60 millivolt, the amount of hydrogen ions which actually reaches the surface of the protein molecule during a certain period of time, and which, therefore, is able to catalyze the reaction, is increased by a factor of 10; or, in other words, the pH, *locally,* on the surface of the protein molecule is decreased by one unit. Similar phenomena occurring on macroscopial and micellar surfaces have been observed and similarly explained by Danielli[41] and by Hartley and Roe.[21]

The high intrinsic affinity of the surface active ions to the proteins requires an explanation. It is likely to have the same cause as the surface activity itself, that is, the squeezing out of the hydrocarbon groups of the surface active ions by the water. We have to assume that the surface active ions are absorbed in such a way that their hydrocarbon portion is in contact with the hydrophobić groups of the protein molecule.[41a] At the same time, the ionic groups of the surface active ion may approach the oppositely charged ionic groups of the protein. A further possibility which merits consideration is that the hydrocarbon parts of the absorbed ions are in mutual contact. The average distance of ionic groups of the same charge in the extended protein chain is about 35 Å or more; the average length of the hydrocarbon portion of the surface active ions is 25 Å. Mutual contact of the hydrocarbon residues of the absorbed ion is, therefore, likely to occur at saturation, especially if the protein chain is not fully extended. As reported above, Lundgren, Elam and O'Connell, as well as Putnam and Neurath, came to the conclusion that if the surface active ions are not in excess, a fraction of the protein molecules is combined with a relatively large amount of surface active ions (e.g., egg albumin with 33% by weight sodium alkyl benzene sulfonate, serum albumin with 22% dodecyl sulfate) and the other fraction is not combined at all. This can possibly be explained by the tendency of the surface active ions to form lipophil islands through mutual contact of their hydrocarbon groups.

DENATURATION, PRECIPITATION AND DISSOLVING OF PROTEINS BY SURFACE ACTIVE IONS

Anson[18] showed that synthetic detergents, anion active as well as cation active, and bile salts denature proteins, such as haemoglobin and egg albumin, at their isoelectric point and keep the denatured protein

[41] Danielli, J. F. Proc. Roy. Soc. London B 122: 155. 1937.
[41a] Dyes are probably adsorbed in the same fashion. Cf. Sheppard, S. E., & A. L. Geddes. J. Chem. Phys. 13: 63. 1945.

in solution. He found that the concentration of detergents required to denature the protein is proportional to the amount of protein present, and concluded that a combination of the protein with the detergent occurs in the process.

Jarisch[42] observed that serum precipitates with soaps and dissolves again in an excess of soap. Matsumura,[43] working in Pauli's laboratory, studied this phenomenon in greater detail and found that purified horse serum is precipitated with 0.01 M sodium oleate and is completely dissolved again with 0.025 M sodium oleate. The interpretation of this phenomenon was obscured by the complicating factor of the hydrolysis of the soap.

Bull and Neurath[44] were the first to mention the precipitating effect of sodium dodecyl sulfate (SDS) on egg albumin and the dispersing effect of the excess of this reagent on the precipitation. Anson[18] observed that detergents in sufficient concentration can prevent the precipitation of denatured protein by trichloroacetic acid, tungstic acid and acid ferric sulfate.

In the following investigations, full advantage has been taken of the fact that synthetic surface active ions possessing strongly acidic or basic groups are not subject to hydrolysis, so that the effect of the pH and the concentration could be observed unobscured. Kuhn and Bielig[45] found that proteins were precipitated by surface active cations only if the former were present as anions. The proteins have been precipitated in weakly alkaline solution, but not in presence of dilute sulfuric acid. The only investigated protein which could not be precipitated in alkaline solution was the strongly basic protein, salmin. A fresh precipitate of egg albumin could be dissolved by an excess of the surface active cation. McMeekin[46] reported that the relative effectiveness of alkyl sulfate and sulfonate, as precipitating agents for proteins, increases in the homologous series until a maximum of precipitate per mole of reagent is reached. Further addition of methylene groups to the molecule is without effect. Benzene ring and branch chain sulfonates are less effective than the corresponding straight chain compounds.

Miller and Andersson[47] reported the precipitating and dispersing effect of SDS on insulin.

[42] Jarisch, A. Pfugers Arch. Physiol. 194: 337. 1932.
[43] Matsumura, S. Kolloid. Z. 32: 173. 1923.
[44] Bull, H. B., & H. J. Neurath. J. Biol. Chem. 118: 163. 1937.
[45] Kuhn, R., & H. J. Bielig. Ber. 73: 1080. 1940.
[46] McMeekin, T. L. Federation Proc. 1 (2): 125. 1942.
[47] Miller, G. L., & K. J. I. Andersson. J. Biol. Chem. 144: 475. 1942.

Putnam and Neurath[34] observed that anionic detergents precipitate proteins only when the latter are in the cationic form. Therefore, it could be generally stated that the surface active ions will precipitate the proteins only when they are oppositely charged to the proteins. However, this is only true in the absence of salt of high concentration (See below).

Schmidt,[48] using the cationic surface active agent, dodecyl dimethyl benzylammonium chloride, and Putnam and Neurath, using the anionic

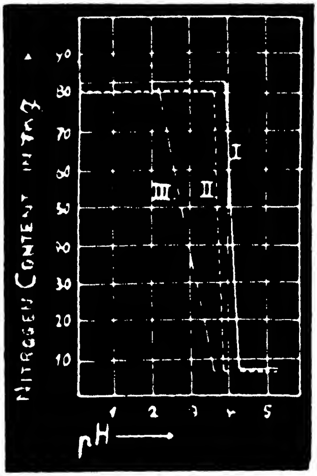


FIGURE 6 Dependence of range of precipitation of protein by surface active cations on the concentration of neutral salt Curve I . 100 ml solution containing 0 558 g. Casein, 0 285 g dimethyl benzyl alkyl ammonium chloride (alkyl C_8-C_{18}), 0.25 M HCl. Curve II: Same + 0.2% NaCl. Curve III Same + 0 4% NaCl. Schmidt, Z physiol Chem **277**: 117. 1942.

surface active agent, SDS, investigated quantitatively the influence of the hydrogen ion concentration on the amount of precipitated protein, at a constant concentration of detergent and protein. The precipitation starts at the isoelectric point of the protein and increases linearly with increasing pH (in presence of the surface active cation) or decreasing pH (in presence of the surface active anion) until all protein present is precipitated, the whole range varying according to conditions from 0.25 to about 2 pH units.

Schmidt found that, by addition of a neutral salt, the pH range of precipitation can be considerably shifted (FIGURE 6). When 0.4% NaCl is present, the precipitation occurs in the pH range 3.5–2, i.e., on

[48] Schmidt, K. H. Z. physiol Chem. **277**: 117. 1942.

the acidic side of the isoelectric point, where the protein and the surface active cation possess the same kind of electric charge.

Putnam and Neurath observed that the amount of surface active agent required to precipitate a certain fraction of the protein at a con-

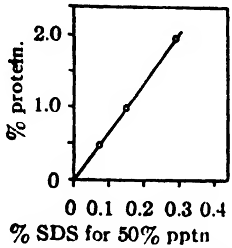


FIGURE 7. Per cent dodecyl sulfate required to precipitate 50% of the serum albumin as a function of the concentration of the protein. Putnam & Neurath J. Am Chem. Soc **66: 692. 1944**

stant pH increases with increasing concentration of the protein present. The linear relationship is indicative of the combination of the protein and the surface active ion (FIGURE 7). The results presented in FIG-

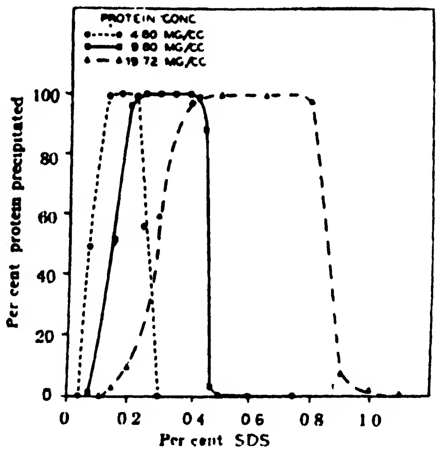


FIGURE 8 Per cent serum albumin precipitated at pH 4 5 in 0.1 N sodium acetate plotted against concentration of sodium dodecyl sulfate from solutions of three different initial protein concentrations. Putnam & Neurath. J. Am Chem. Soc. **66**: 692. 1944.

URE 8 show three distinct regions of the interaction of protein with surface active agent: (a) that of protein excess where the protein is incompletely precipitated; (b) the equivalence zone where complete

precipitation is achieved; and (c) the region of detergent excess in which the precipitate may be partly or completely dispersed. The zone of complete precipitation corresponds to a weight ratio, 5:1 to 2.5:1 protein-detergent, which is equivalent to about from one half to the total of the acid-binding capacity of the protein. If the weight ratio reaches the value 1:1, the protein is again completely dispersed.

The precipitating action of surface active ions on proteins seems to follow the same mechanism as a great many others of the well-known precipitating reactions of the proteins. This mechanism can be described as elimination of the free electric charge of highly ionized protein through interaction with counter-ions. Accordingly, these reactions are carried out by increasing the total charge, that is, by increasing the amount of adsorbed hydrogen ions or hydroxyl ions, and by adding, simultaneously, counter-ions which possess high affinity for the protein.[35] Precipitation with heavy metal salts is carried out in alkaline solution; precipitation with complex acids, acid dyes, ferrocyanides and picrates is carried out in acid solution. It is surprising that the stability of the proteins is not at a minimum at the isoelectric point, but rather in the ionized form. This can be understood on the basis of the theory of zwitter-ionization of the proteins. According to this, at the isoelectric point (or rather at the iso-ionic point), the ionic groups of the protein, that is, the amino and carboxylic groups, are ionized. To a certain degree, an intermolecular neutralization takes place, yet the protein retains a sufficient amount of ionic charge to effect hydration. As acid is added, the ionization of the carboxylic groups is suppressed and the protein acquires a positive charge. When, by association with anions, the ionic charge of the ammonium groups is reduced, the residual hydration of the proteins is eliminated and consequently the stability of the protein is destroyed.

The dispersing effect of excess detergent, likewise, can be easily understood by considering the electric charge of the protein. When, after neutralization of the free electric charge of the protein, an additional amount of free surface active ions is adsorbed, the protein acquires an increasing electric charge of the same sign as the surface active ion and, therefore, its hydration is increased. This dispersing effect seems to be characteristic for the interaction of the proteins with surface active ions and does not seem to occur with other types of precipitating agents. It may be due to the adsorption of the surface active ions in a second layer on the top of the first layer which covers the surface of the protein molecule. The outside of the second layer is likely to be composed of the ionizing groups of the surface active

ions, the hydrocarbon portion in the second layer facing the hydrocarbon portion of the first layer.

Schmidt, as well as Putnam and Neurath, observed that precipitation of protein by surface active ions is promoted by the presence of a neutral salt. Recently, Pankhurst and Smith[49] investigated the effect of ammonium nitrate on the separation of an oily phase from a gelatin solution in the presence of varying amounts of SDS. If the pH was kept constant at 5.3, which is slightly on the alkaline side, no separation occurs until the salt concentration is above 0.35 M. At higher salt concentration, the separation increased with increasing concentration of SDS until a maximum had been reached. Further addition of SDS above this maximum decreased the amount separated. The SDS concentration at maximum separation was practically independent of the ammonium nitrate concentration as the latter varied between 0.45 M and 2.0 M. However, it was proportional to the gelatin concentration corresponding to the weight ratio of about 3:1 detergent-gelatin. This weight ratio decreased with decreasing pH and at pH 1 it amounted to only about 1:1. The effect of neutral salt on the precipitation and separation of the protein-detergent complexes can be explained as the salting out of the combination protein-surface active ion. The ionization of the complex is probably depressed at the high concentration of the electrolytes in the solution.

The interaction of surface active ions and proteins found an interesting application in the study of the formation of artificial protein fibers. On the one hand, the denaturation of corpuscular proteins was thought to be accompanied by their transformation into fibrous proteins,[50] a transformation due to the unfolding or uncoiling of the native protein molecules. On the other hand, as Anson has shown, detergents are excellent denaturing agents. Since the degrading action of the detergents on the protein molecules is relatively mild, it is not surprising that, as Lundgren and O'Connell[51] found, synthetic detergents can be advantageously employed as solvents in the preparation of protein dispersions for the manufacture of protein fibers. Fibers, so prepared from egg albumin, show the X-ray pattern of linear polypeptide chains (B-keratin configuration).[52]

The precipitation of proteins with surface active ions promises to be a useful method for the characterization of proteins and the separation of their mixtures.[48] However, it can not be used for the purification

[49] Pankhurst, K. G. A., & R. C. M. Smith. Trans. Faraday Soc. 40: 565. 1944.
[50] Astbury, W. T., S. Dickinson & K. Bailey. J. Biochem. 29: 2351. 1935.
[51] Lundgren, H. P., & R. A. O'Connell. Ind. Eng. Chem. 36: 370. 1944.
[52] Palmer, K. J., & J. A. Galvin. J. Am. Chem. Soc. 65: 2187. 1943.

of native proteins, since the amount of surface active ions required for precipitation is higher than the amount necessary to cause denaturation.

EFFECT OF SURFACE ACTIVE AGENTS ON ENZYMES, SIMPLEXES AND VIRUSES

It has been long known that soaps inhibit the proteolytic activity of trypsin. This was recently confirmed by Peck[53] using crystalline trypsin and pure soap. Marron and Moreland[54] found very little effect of Alphasol LA, an anionic detergent on urease. Keilin and Hartree[55] observed that cytochrom C is reversibly changed by sodium dodecyl sulfate. The change apparently affects the linkage of the heme group to the protein since the absorption spectrum is modified. Kuhn and Bielig[45] reported that catalase is precipitated and completely inactivated in weakly alkaline medium by 0.33% of cationic surface active agents. In neutral or weakly acid solution, no precipitation or inactivation takes place. Shoch and Fogelson[56] found that pepsin is completely inhibited by 0.5% sodium lauryl sulfate. Following this, experimental work has been carried out by various investigators[57,11,58,12,59] concerning the effect of sodium lauryl sulfate on the peptic activity of the gastric contents and on the healing of gastric ulcer in man. However, in respect to the latter, as yet, no unequivocal conclusion has been reached. Miller and Andersson[47] have shown by ultracentrifugal and diffusion measurements that insulin, in 1% solution, is split into smaller molecules when treated with 2% sodium lauryl sulfate.

Bile salts have been used to extract the photosensitive pigment protein of the dye,[60] and to extract a chlorophyll compound from the chloroplast of spinach.[61] Anson[18] suggested the use of synthetic detergents for these purposes. Smith[62] investigated the action of sodium dodecyl sulfate (SDS) on the chlorophyll protein compound of the spinach leaf and found that SDS clarified the alkaline extract with much greater effectiveness than digitonin or bile salts. Magnesium may be eliminated by this treatment. However, the chlorophyll or the phaeophytin remains attached to the protein. Smith and Pickels[63] investigated the action of surface active agents on the chloroplast with the ultra-

[53] **Peck, R. L.** J. Am. Chem. Soc. **64:** 487. 1942.
[54] **Marron, T. U., & F. B. Moreland.** Enzymologya **6:** 225. 1939.
[55] **Keilin, D., & E. F. Hartree.** Nature **145:** 934. 1940.
[56] **Shoch, D., & S. J. Fogelson.** Proc. Soc. Exp. Biol. Med. **50:** 304. 1942.
[57] **Steigmann, F., & A. R. Marks.** Proc. Soc. Exp. Biol. Med. **54:** 25. 1943.
[58] **Kirsner, J. B., & H. H. Spitzer.** Gastroenterology **2:** 348. 1944.
[59] **Kirsner, J. B., & I. A. Wolf.** Gastroenterology **2:** 270. 1944.
[60] **Kuhne, W.** F. C. W. Vogel **3:** 264. 1879.
[61] **Smith, E. L.** Science **88:** 107. 1938.
[62] **Smith, E. L.** Science **91:** 199. 1940; J. Gen. Physiol. **24:** 565. 1941; J. Gen. Physiol **24:** 583. 1941.
[63] **Smith, E. L., & E. G. Pickels.** J. Gen. Physiol. **24:** 753. 1941.

centrifuge and found that digitonin, bile salts and sodium desoxycholate split the chlorophyll from the protein, but the protein remains of high molecular weight. On the other hand, SDS does not split off the prosthetic group from the protein but decomposes the protein into smaller units. Kuhn and Bielig[45] observed the cleavage of chloroplast by surface active cations such as lauryl dimethyl benzyl ammonium bromide and dodecyl dimethyl sulfonium iodide. In contrast to these surface active cations, sodium palmitate and alkyl sulfate did not split off the chlorophyll from the chloroplast of hydrangea leaves. Surface active cations proved also to be useful in the extraction of carotene from the juice of carrots. Chromoproteides, with the exception of ovoverdine, could not be dissociated with surface active cations.

Several investigations were devoted to the action of surface active agents on viruses. Bawden and Pirie[64] observed that preparations of potato virus X and bushy stunt virus are destroyed by incubation with 0.3% SDS. According to Sreenivasaya and Pirie,[65] tobacco mosaic virus disintegrates when treated with SDS, sodium cetyl sulfate and sodium N-oleyl-N-methyl tauride, the first being the most effective. The virus preparations not only lose their infectiousness by this treatment, but at the same time, the nucleic acid separates from the protein component to which it was attached. Loss of sedimentability of the protein component in the centrifuge indicates that the protein splits into relatively small particles. However, comparatively high concentration, for instance, 1% SDS is required to inactivate and disintegrate this virus. Pfankuch and Kausche[66] have shown that tobacco mosaic virus and potato virus X are precipitated and inactivated by surface active cations In contrast to the action of the surface active anions, no cleavage into nucleic acid or into smaller protein units occurs. The inactivation and the increase of the particle size determined by measurement of the turbidity were, in most cases, parallel. However, some of the investigated compounds inactivated the virus without changing its state of dispersion.

In studying the inactivation of the virus of epidermic influenza by soaps, Stock and Francis[67] found that the sodium salts of oleic, linoleic and linolenic acid were the most effective. In comparison with them, SDS was only moderately effective.

Krueger[68] and collaborators observed that alkyl dimethyl benzyl ammonium chloride in a dilution of 1:10,000 (alkyl: C_8–C_{18}) fails to

---

[64] **Bawden, F. C., & N. W. Pirie.** Brit. J. Exp. Path. **19:** 66. 1938.
[65] **Sreenivasaya, M., & N. W. Pirie.** Biochem. J. London **32:** 1707. 1938.
[66] **Pfankuch, E., & G. A. Kausche.** Biochem. Z. **312:** 72. 1942.
[67] **Stock, C. C., & T. J. Francis.** Exp. Med. **71:** 661. 1940.
[68] **Krueger, A. P., & Unit Personnel.** U. S. Naval Med. Bull. **40:** 622: 1942.

inactivate influenza virus preparations in one hour of exposure. Soap, in the dilution of 1:1,000 and 1:100, was effective. Because this virus proved to be more resistant to the surface active cations than bacteria, it was possible to free influenza virus contained in throat washings, egg fluids, or ground mouse lung menstrua from adventitious bacterial contamination by treatment with a 1:20,000 dilution of the cationic germicide.[69]

Soaps and detergents cause lysis of cells, particularly of the red blood cells. The cytolytic action of a series of surface active anions has shown close parallelism with their surface activity.[70]

MASS RELATION IN THE INTERACTION OF SURFACE ACTIVE IONS WITH PROTEINS

Since surface active ions are strongly adsorbed from their aqueous solutions by proteins, the amount of surface active ions which is necessary to cause a certain effect should be generally related to the amount

TABLE 1
WEIGHT RATIO IN THE INTERACTION OF SURFACE ACTIVE IONS WITH PROTEINS

Investigator	Protein	Effect	Conc. of Protein Per cent	Protein: Detergent
Sreenivasaya & Pirie[65]	Virus	Splitting	0.5	1:1
Anson[18]	Methemoglobin	Denaturation	0.1	5:1
Kuhn & Bielig[45]	Egg albumin	Denaturation	0.6	100:1
Kuhn & Bielig[45]	Egg albumin	Precipitation	0.6	2:1
Kuhn & Bielig[45]	Egg albumin	Dissolving	0.6	1:1
Miller & Andersson[47]	Insulin	Dispersion	1	1:2
Miller & Andersson[47]	Serum albumin	Combination	1	2:1
Pfankuch & Kausche[66]	Virus	Precipitation	0.04	4:1
Pfankuch & Kausche[66]	Virus	Turbidity	0.04	40:1
Pfankuch & Kausche[66]	Egg albumin	Dissolving	0.04	1:1
Schmidt[48]	Casein	Precipitation	0.5-1	2:1
Lundgren, Elam, O'Connell[37]	Egg albumin	Combination	1	3:1
Putnam & Neurath[34]	Serum albumin	Precipitation	1	4:1
Putnam & Neurath[34]	Serum albumin	Dissolving	1	1:1
Putnam & Neurath[38]	Serum albumin	Combination	1	4.5:1

of protein present rather than to the volume of the system. Only if the protein is present in a very high dilution, does the concentration of the surface active ions become the significant quantity. TABLE 1 presents a few data of the required weight ratio of protein to surface

[69] Krueger, A. P., & U. S. Naval Unit. Science 96: 543. 1942.
[70] Hober, R., & J. Hober. J. Gen. Physiol. 25: 705. 1942.

active agent, calculated approximately on the basis of available experimental material. It is remarkable that, with the exception of the turbidity and the denaturation, the various phenomena required the presence of surface active agents in an amount of at least 20% of the weight of the protein. Since the average molecular weight of the surface active agents is about 400, it means that, for every 2,000 grams of protein, at least one mole of surface active ion was present in the system where the interaction took place.

EFFECT OF SURFACE ACTIVE AGENTS ON BACTERIA

It has been known for a long time that ordinary soaps have certain disinfectant properties. The development of surface active germicides is relatively new. Hartmann and Kaegi[6] reported in 1928 the finding of Doerr, that their newly synthesized surface active cations were germicidal. These surface active cations were prepared by condensation of higher fatty acids with unsymmetrical diethyl ethylene diamine and also by alkylation of the resultant N-acyl amidoethyl-N'diethyl amines. However, this report, at first, attracted little attention and so remained without practical importance. In 1933, Domagk observed the extraordinarily high germicidal power of dodecyl amine hydrochloride, which, for the determination of its antibacterial properties, has been submitted to him by the writer.[71] The dilution of this compound, killing pathogenic germs of the type of *Staphylococcus aureus* and *Eberthella typhosa* was approximately 1:10,000. Expressed in terms of the F. D. A. method, this means a phenol coefficient of about 120–150. This observation immediately suggested the investigation of the derivatives of dodecyl amine which resulted after a few months in the commercial development of a new germicide, dodecyl dimethyl benzyl ammonium chloride. The phenol coefficient of this compound is slightly more than twice that of the primary amine, the killing dilution for the common pathogenic germs being between 1:20,000 and 1:30,000.

Domagk published a report on the bacteriological properties of the new germicide in 1935.[19] Since then, an ever increasing number of germicidal cations including ammonium, phosphonium, sulfonium and arsonium ions with different types of substituents have been recorded in the scientific and patent literature.

[71] This sample was submitted in November 1932 by the writer for blank test along with another sample containing a mercury complex of dodecyl amine and excess dodecyl amine hydrochloride as solubilizing agent. Domagk published an incomplete account[19] of the development of the new class of germicides by failing to mention this.

Cowles,[72] as well as Birkeland and Steinhaus,[73] found that the higher sodium alkyl sulfates exert a pronounced inhibiting effect upon Gram-positive but not upon Gram-negative organisms. Previously, Katz and Lipsitz[74] found that sodium di-(secondary) butyl naphthalene sulfonate is bactericidal against *Mycobacterium smegmatis;* and Baylis and Halvorson[75] reported that soaps as well as alkyl sulfates are germicidal against pneumococci (*vide* also[76]).

Baker, Harrison and Miller[77] investigated the inhibiting action of surface active cations and anions on the metabolism of bacteria. They found that the surface active cations inhibit the metabolism of Gram-positive and Gram-negative micro-organisms in the same degree. On the other hand, the anions were found to be effective only against Gram-positive organisms. The inhibiting action of the cations increased with increasing pH, and that of the anions decreased. In a subsequent paper[78] these authors found that the germicidal behavior of the surface active agents corresponded closely to their inhibiting action against the same organism. Dubos[79] reported likewise a difference in activity between the anionic and cationic detergents against Gram-positive and Gram-negative species.

According to Gershenfeld and Perlstein,[80] sodium di-octyl sulfo succinate, an anionic surface active agent, is effective against *Staphylococcus aureus* at low pH values. For instance, the killing dilutions for five minutes exposure at pH 6 was 1:4,000, and at pH 4, 1:40,000. Gershenfeld and Milanick[81] observed that, when the acidity of the solution was high enough, surface active anions were germicidal even against Gram-negative organisms. Scales and Kemp[82] likewise observed the germicidal effectiveness of surface active anions against Gram-positive and Gram-negative organisms at pH 4.

Soaps and anionic detergents are extensively used in dissolving phenols. Their synergistic effect,[83] which was definitely proven recently by Ordal, Wilson and Borg,[84, 85] working with buffer solutions, has often been obscured by variation of the hydrogen ion concentration. Descriptive bulletins and patents also claim synergistic effect of ionic de-

[72] **Cowles, P. B.** Yale J. Biol. and Med. **11**: 33. 1938.
[73] **Birkeland, J. M., & S. A. Steinhaus.** Proc. Exp. Biol and Med. **40**: 86. 1939
[74] **Katz, J., & A. Lipsitz.** (I) J. Bact. **30**: 419. 1935; (II) J Bact. **33**: 479. 1937.
[75] **Baylis, M., & H. O. Halvorson.** J. Bact. **29**: 9. 1935.
[76] **Baylis, M.** J. Lab. Clin. Med. **22**: 700. 1936.
[77] **Baker, Z., R. W. Harrison & B. F. Miller.** J. Exp. Med. **73**: 249. 1940.
[78] **Baker, Z., R. W. Harrison & B. F. Miller.** J. Exp. Med **74**: 621. 1941.
[79] **Dubos, R. J.** Ann Rev. Biochem **11**: 659. 1942.
[80] **Gershenfeld, L., & D. Perlstein.** Am J. Pharm. **113**: 237. 1941
[81] **Gershenfeld, L., & V. E. Milanick.** Am. J. Pharm. **113**: 306. 1941
[82] **Scales, F. M., & M. Kemp.** Proc. Intern. Assoc. Milk Dealers **33**: 491. 1941.
[83] **Cade. A. R.** Soap **11**(9): 27. 1935.
[84] **Ordal, E. J., J. L. Wilson & A. F. Borg.** J. Bact. **42**: 117. 1941.
[85] **Ordal, E. J., & F. Deromedi.** J. Bact. **45**: 293. 1943.

tergents with other types of disinfectant such as aryl sulfonic chloroamide[86] or mercuric salts.[87]

Petroff and Schain[88] have carried out phenol coefficient determinations on mixtures of a large number of surface active compounds with various germicides. The combination of N-N′-dichloro-azo-dicarbonamidine (Schmelkes[89]) with sodium 7-ethyl-2-methyl-undecyl-4-sulfate (tetradecyl sulfate) proved to be extraordinarily effective. This mixture was used for therapeutic purposes, namely, for irrigation in the treatment of empyema. According to Kintz,[90] the same combination showed good results in the treatment of infectious wounds of the soft tissues of the face (*vide* also Salle[91]). Further investigations on the influence of surface active agents on various antiseptics have been carried out by Gershenfeld and Perlstein[80] and by Fisher [92]

Several attempts have been made to establish a correlation between chemical structure and antibacterial action, but none of the attempted correlations can claim general validity. One of the first systematic studies in this field was made by Stanley, Coleman, Greer, Sacks and Adams.[93] A very large number of sodium salts of synthetic organic acids have been investigated for their bactericidal efficacy against *Mycobacterium leprae* and *Mycobacterium tuberculosis*. The experimental results definitely proved that one of the important factors governing the bactericidal effectiveness is the molecular weight, with the maximum appearing ordinarily in molecules of 15 to 18 carbon atoms. It was found that the bactericidal acids were also good surface tension depressors. Decrease in the molecular weight of the acids below 256 caused a drop in bactericidal effectiveness, and a parallel decrease in surface tension. Increasing the molecular weight above 256 caused a decrease in bactericidal effectiveness without the corresponding lowering of the surface tension. This, surface activity is also but one of two or more factors, which, in the opinion of Stanley, Adams and collaborators, may be responsible for the bactericidal action of the investigated aliphatic acids.

In the course of this work, Adams and collaborators prepared a series of cyclohexyl substituted amines, especially tertiary amines of the formula:

$$C_6H_{11}(CH_2)_xN(C_2H_5)_2 . HCl$$

[86] **Heyden, A. G.** French Patent No **731,395.** 1935.
[87] **Valko, E. I., & J. Nusslein.** German Patent No. **657,116.** 1932
[88] **Petroff, S. A., & P. Schain.** Quarterly Bulletin of Sea View Hospital **372.** 1940
[89] **Schmelkes, F. C.** U. S. Patent **1,958,370.** 1934.
[90] **Kintz, F. E.** Military Surgeon **89:** 61. 1941.
[91] **Salle, A.** J. Proc. Exp Biol. Med. **49:** 141 1941.
[92] **Fisher, C. V.** Am J. Pub. Health **32:** 389 1942.
[93] **Stanley, W. M., G. H. Coleman, C. M. Greer, J. Sacks & R. Adams.** J. Pharmacol. and Exp. Ther. **45:** 121. 1932.

x ranging from 0 to 6. In these compounds as with the acids, the molecules of 16–18 carbon atoms possessed the greatest bactericidal action. The amines, however, were not as effective as the acids. Thus, Adams and collaborators observed, prior to Domagk, the bactericidal effect of high molecular aliphatic amines in a special field, namely, against acid fast bacteria, where these, however, did not prove superior to surface active anions.

The observation of a comparatively sharp peak in the curve representing the dependency of antibacterial effect on the chain length in the homologous series was confirmed by a number of investigators: In the group of alkyl dimethyl sulfonium iodides,[94] of dialkyl benzotriazolium chlorides,[95] of alkyl triethyl phosphonium and arsonium compounds, and alkyl diethyl benzyl phosphonium and arsonium compounds,[96] of the alkyl pyridinium and alkyl picolinium compounds,[97] of P-alkyl phenoxy ethoxyethyl N-dimethyl N-benzyl ammonium chlorides,[98] of the fatty acid esters of the hydroxyethylamide of pyridinbetaine,[99] (*vide* also [100-104]). In the presence of only one higher hydrocarbon group in the molecule, the optimum was reached almost always at a content of 12, 14, or 16 carbon atoms in this group. However, there was some indication that the influence of the chain length of the high molecular radical and that of the nature of the other substituents of the nitrogen atom are interdependent. This was clearly shown, recently, by a comparison of the germicidal efficiency of primary amines and of trimethyl and dimethyl benzyl ammonium compounds.[105]

MECHANISM OF THE ANTIBACTERIAL EFFECT

Probably, the initial process in the action of surface active ions on bacteria is their reversible adsorption by, or, combination with the bacteria.[106, 107, 108] This conception can be traced back to a certain ex-

[94] **Kuhn, R., & O. Dann.** Ber. **73:** 1092. 1940.
[95] **Kuhn, R., & O. Westphal.** Ber. **73:** 1105. 1940.
[96] **Jerchel, D.** Ber. **76:** 600. 1943.
[97] **Kolloff, H. G., A. P. Wyss, R. E. Himelick & F. Mantele.** J. Am. Pharm. Assoc. **31:** 51. 1942.
[98] **Rawlins, A. L., L. A. Sweet & D. A. Joslyn.** J. Am. Pharm. Assoc. **32.** 1943.
[99] **Epstein, A. K., B. R. Harris & M. Katzman.** Proc Soc. Exp. Med. **53:** 238. 1943.
[100] **Leffler, M. T., & H. R. Volwiler.** Meeting Am. Chem. Soc. Boston. 1939.
[101] **Shelton, R. S., M. G. Van Campen & L. Nisonger.** Meeting Am. Chem. Soc. Boston. 1939.
[102] **Shelton, R. S., M. G. Van Campen, C. H. Tilford & L. Nisonger.** Meeting Am. Chem. Soc. Cincinnati. 1940.
[103] **Niederl, J. B., & H. Weingarten.** J. Am. Chem. Soc. **63:** 3534. 1941.
[104] **Eldred, F. R., & J. B. Niederl.** Meeting Am. Chem. Soc. Atlantic City. 1941.
[105] **Valko, E. I., & A. S. DuBois.** J. Bact. **50:** 48. 1945.
[106] **Valko, E. I., & A. S. DuBois.** 104th Meeting Am. Chem. Soc. Buffalo. 1942.
[107] **Valko, E. I., & A. S. DuBois.** J. Bact. **47:** 15. 1944.
[108] **Albert, A.** Lancet [illegible]. 1942.

tent to Ehrlich's assumption that the basis for the antibacterial action, is a combination between the antibacterial agent and some of the nutriceptors of the bacteria. Simon and Wood[109] adopted this conception as a basis for the antibacterial effect of basic dyes and assumed the existence of acidic groups in the structure of the bacterial organism with which the ammonium groups of the inhibiting dyes would unite. The cell dies, according to them, because a sufficient number of its nutriceptors have been thrown out of action, bringing about its starvation or inability to multiply. Stearn and Stearn[110] accepted the same view in respect to the action of toxic cations generally, including those of the heavy metal salts and basic dyes as well as the cations of the substituted aniline type. They assumed that the electrochemical equivalent weight of the bacteria is essentially equal to that of their proteins. McCalla[111] measured the adsorption of hydrogen ions, cationic dyes, mercuric and calcium ions by bacteria and found that the maximum amount bound in one equivalent per 1200–2000 grams of bacteria regardless of the nature of the ion.

Generally, dye anions and cations permeate the bacteria in a short time, (acid fast bacteria, are, of course, excepted). There is no reason to assume that surface active ions will not penetrate into the bacteria Indeed, Kuhn and Jerchel[112] proved that surface active cations penetrate into living cells. Tetrazolium compounds are colorless, but can be reduced to red colored compounds. Yeast, treated with a nutrient medium containing tetrazolium salts, became red colored when the compound was adsorbed and then reduced in the cells Pathogenic germs exhibited the same behavior. Therefore, it does not seem warranted to regard the interaction of a hypothetical bacterial membrane with surface active ions to be important in the first phase of the antibacterial action.

Valko and DuBois[107] demonstrated that, like the action of mercuric chloride (Engelhardt[113]), the so-called killing action of surface active cations on bacteria can be reversed under certain conditions by detoxication of the bacteria with a high molecular anion. Sodium dodecyl sulfate was used as detoxicating agent. The fact that mere dilution with water does not restore viability of the bacteria was attributed to the slow establishment of the desorption equilibrium. These authors have shown that the addition of relatively harmless but adsorbable ca-

[109] **Simon, C. E., & M. A. Wood.** Am. J Med Sci **147** (New Series) · 524
[110] **Stearn, A. E., & E. W. Stearn.** Univ. Missouri Studies **3**(2). 1928
[111] **McCalla, T. M.** J. Bact. **41:** 775. 1941; J. Bact. **40:** 23. 1940.
[112] **Kuhn, R., & D. Jerchel.** Ber. **74:** 941, 949. 1941.
[113] **Engelhardt, H.** Desinfektion **7:** 81. 1922.

tions exerts a protective action on the bacteria against more toxic anions and pointed out that the dependency of the antibacterial action on the pH appears to be due to the dependence of the cationic and anionic adsorption on the pH. The dependency on the pH fits into the general scheme of protective action of adsorbable but relatively harmless ions (hydrogen ions and hydroxyl ions) which have the same sign as the antibacterial ion. This dependence can be deduced from the electrostatic interaction in accordance with Equations (1)–(5) of this paper.

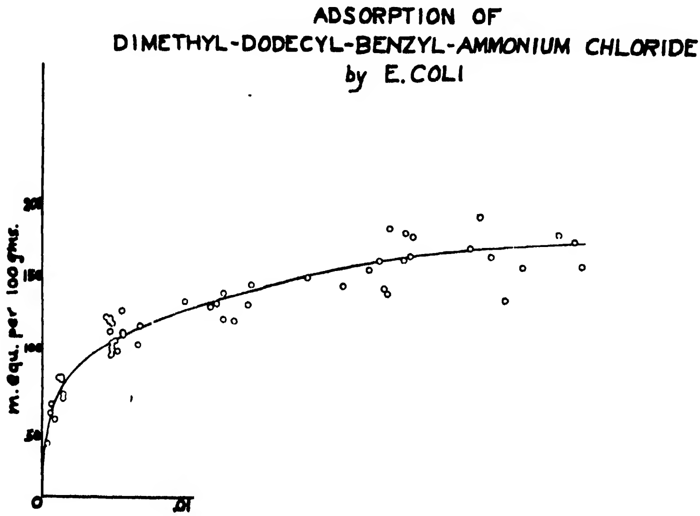


FIGURE 9. Adsorption of dimethyl dodecyl benzyl ammonium chloride (m. eqv. per 100 grams of bacteria) by *Escherichia coli* as a function of the molarity of the surface active ion in the supernatant liquid. Valko & Dibblee. 1944.

A distinction should be made between the apparent toxicity and specific toxicity of the ions to the bacteria. Apparent toxicity is related to the concentration of the killing solution; specific toxicity is defined in relation to unit adsorbed ions. Accordingly, apparent toxicity is the product of two factors; namely, adsorbability and specific toxicity. It would be desirable to resolve experimentally the germicidal efficacy into these two factors. To do this, one has to measure the adsorption of toxic ions by bacteria. Valko and Dibblee[114] determined the adsorption equilibrium of *Escherichia coli* with dimethyl dodecyl benzyl ammonium chloride (FIGURE 9). The surface active agent added to a

[114] Valko, E. I., & D. Dibblee. Not yet published. 1944.

suspension of bacteria and, after centrifuging, the supernatant liquid was analyzed for its content of surface active cations. The results show that 100 grams of bacteria (dry weight) combine with a maximum of about 60 grams of surface active cations This corresponds to an equivalent weight of the bacteria of about 600, which is approximately the same as the equivalent weight of many animal proteins. With *Escherichia coli,* which is relatively resistant to this germicide, the bacteria are nearly saturated with the germicidal ion, if the latter is present in a dilution which kills in five minutes. However, the adsorption was measured only after the bacteria had been in contact with the solution for a considerably longer period.

The reversible ionic adsorption has to be considered only as a first process in the mechanism of the antibacterial action. This reversal can be carried out after the bacteria has been exposed to the action of the toxic cations only for a relatively short time. If the primary process is not reversed soon, some other processes set in and finally cause the death of the bacteria. Little is known about the course of these events. According to Stearn and Stearn,[110] the main constituent of bacterial cells is a combination of protein with nucleo-proteins and lipo-proteins. This combination is essentially an electrostatic one and the adsorption of ions causes a disturbance of this equilibrium. Kuhn and Bielig[45] expressed the opinion, that the killing action of surface active cations is probably due to the interaction of the ions with anionic proteins and the essential simplexes of the bacterial cells. Baker, Harrison, and Miller[115] have shown that the antibacterial effect of both cationic and anionic detergents can be inhibited by simultaneous addition of appropriate amount of phospholipides. Bacteria treated with phospholipides and washed with water retained a resistance against the toxic surface active ions. Baker, Harrison and Miller believe that the most reasonable working hypothesis to explain the rapid action of surface active ions on bacterial metabolism is probably the one based on a twofold action: First, disorganization of the cell membrane by virtue of the great surface activity of these compounds; and, second, denaturation of certain essential proteins.

More experimental data are required before it can be definitely stated whether one or more of the suggested mechanisms, namely, denaturation and precipitation of proteins; cleavage and inactivation of simplexes, particularly enzymes; lysis of cells (*vide* Hotchkiss[116]) play an

[115] **Baker, Z., R. W. Harrison & B. F. Miller.** J Exp Med 74: 611 1941
[116] **Hotchkiss, R. D.** Ann N Y. Acad. Sci. 46(6) 480-492 1946.

important part in the irreversible action of surface active ions on bacteria.

The affinity of the surface active ions to proteins, which is probably the cause of their antibacterial action, represents at the same time a certain disadvantage in their germicidal performance since it causes a decrease of their effectiveness in the presence of other proteins. Klarmann[117] showed that the presence of serum or blood strongly reduces the antibacterial action of surface active cations. (*Vide* also Mallman[118].) Nevertheless, their effectiveness remains high enough to indicate that the surface active ions possess a selectivity, that is, a preference for the bacteria. Indeed, when tested for relative destruction of living tissue and pathogenic germs, the toxicity index is lower for the best germicidal cations than for the other types of germicides investigated.[119]

Non-ionic surface active molecules, as far as investigated, were found to be less effective in the interaction with proteins or bacteria than the surface active ions. This is easily understood, if the action is based on an electrostatic attraction. It would be difficult to explain if the ionic reagent should prove more effective than the non-ionic, even under such conditions that the combination of the active ions with the protein is opposed by electric forces. Definite proof for this is still lacking, especially in view of the zwitter-ionic structure and polarizability of the proteins.

SUMMARY

The biochemical effects of surface active agents may be at least partly explained as follows: Surface active ions possess extraordinarily strong intrinsic affinity for proteins, and, therefore, combine readily with them. This combination causes disturbance of the intermolecular structure of proteins by upsetting the balance of the electrostatic forces, as well as that of the non-Coulombic cohesion in the molecule. At the same time, the interaction of the proteins with the solvent molecules may undergo profound changes and, as a further consequence, the bonds between the components of the conjugated proteins may be disrupted. Denaturation and unfolding of the protein molecules, inactivation of enzymes, viruses and bacteria are the manifest results of these processes.

[117] Klarmann, E. G., & E. Wright. Soap 20(2): 103. 1944.
[118] Mallman, W. L. Soap 20(8): 101. 1944.
[119] Welch, H., & C. M. Brewer. J. Immun. 43: 25. 1942.

THE NATURE OF THE BACTERICIDAL ACTION OF SURFACE ACTIVE AGENTS

By Rollin D. Hotchkiss

Lt. Comdr., Hospital Corps, USNR. From the United States Navy Research Unit at the Rockefeller Institute Hospital, New York, N. Y.*

BACTERICIDAL ACTIVITY

Among the chemical substances having antiseptic action, the surface active agents appear to represent a well-defined class distinct from, say, the dyes, the oxidizing agents, and the heavy metal compounds. Many substances having chemically little in common except a surface activity and some modification of the hydrophobic-hydrophilic structure which is responsible therefor, are strongly bactericidal. Appropriate cationic agents are able, in dilute solution (10^{-4} to 10^{-5} by weight), to kill organisms from a large variety of bacterial species, and many anionic agents are effective upon representatives of a more restricted group, notably the Gram-positive species. Agents with non-ionizing hydrophilic groups are not commonly bactericidal. Between pH 5 and 8, cationic agents are more effective in alkaline environments, and anionic agents more effective in acid environments.

Few other generalizations may be made that cover the diverse agents toxic for diverse microorganisms. It is one of the purposes of this review to propose some further generalizations that may be reasonably offered at this time.

In particular, we are quite unable to say why, of the vast number of synthetic surface active agents available, certain ones are highly lethal to bacteria and others are much less so. No unique active chemical groups have been recognized. Indeed, two active substances may bear dissimilar groups, and yet other substances bearing these groups may be relatively inactive. Nevertheless, within a homologous series, there is a tendency for an optimally bactericidal substance to exist, which is often close in the series to the substance having a maximal effect on surface tension at the air-water interface. Yet surface tension depression alone, or wetting action, or detergency, or dispersing, or emulsifying ability, as ordinarily measured, does not appear to indicate whether high bactericidal activity will be found within a given homologous series or not. It may, perhaps, be said that these measurements, among

* The views and opinions expressed in this article are those of the writer and are not to be construed as reflecting the official views or opinions of the Department of the Navy.

other things, are all indirect measures of surface activity at arbitrarily chosen interfaces. Thus, ordinary surface tension would be, in part, a measure of the tendency of the hydrophobic elements of the structure to escape from the water phase, but could not include any measure of the avidity with which they would escape from the water, in order to become oriented upon a specific surface, say the surface of a particular species of bacterial cell. Since we know less, either qualitatively or quantitatively, about the nature of cell surfaces than about surfaces of dye molecules or of protein or textile fibers, it is natural that how chemical structure is related to bactericidal activity is even less clear than how it is related to detergency, wetting power, and the like.

Specifically, bactericidal activity is found with straight chain aliphatic, with aromatic and heterocyclic quaternary ammonium compounds, and also with primary amines, sulfonium bases, phenols, carboxylic acids, alkyl and aryl sulfates, and sulfonates, simple and substituted.

INHIBITION OF METABOLISM

What is ordinarily observed is that bacteria suspended in solutions containing small amounts of the active substances, if affected, soon lose the ability to grow in media in which they would normally grow and multiply profusely. To know something about the mechanism of this bactericidal action, we must discover other processes which accompany the death of the cell and are, in effect, irreversible. Very few studies have included observations of phenomena other than growth in treated cells.

Baker, Harrison, and Miller have examined the effect of ten cationic and ten anionic detergents upon the metabolic properties of bacteria.[1] In general, respiratory and glycolytic activity in six Gram-positive and six Gram-negative bacterial species were markedly cut down, the cationic agents being more inhibitory, and the anionic ones affecting mainly some of the Gram-positive bacteria. In a later work,[2] the same authors showed that detergents killed bacteria under the same conditions as had been used in examining the metabolic effects. For the ten detergents and five bacterial cultures common to both studies, it was concluded that depression of metabolism was roughly parallel with killing of the bacteria. However, the regularity of this relation was considerably impaired by the frequent observation that a concentration of detergent might markedly inhibit bacterial respiration, yet, under the

[1] Baker, Z., R. W. Harrison & B. F. Miller. J. Exp. Med. 73: 249 1941.
[2] Baker, Z., R. W. Harrison & B. F. Miller. J. Exp. Med. 74: 611. 1941.

same conditions, appear to be without killing power. This discrepancy may be attributed to the unfortunate choice of a modified phenol coefficient test for judging bactericidal activity. The test was capable of indicating only two results, 0% or 100% killing. For example, from the subculture inoculum of 6×10^6 cells used, full growth would almost certainly have been observed equally well after 24 or 48 hours incubation whether 6×10^6 or only 6 cells remained alive after the treatment. Obviously, if 99.9999% killing gives the same end result as no killing, it is going to be difficult to correlate the results of the growth test with the essentially quantitative data on metabolic inhibition. For example, suppose one finds the metabolism is only 5% of normal after detergent is added. With the all or none growth test, one cannot decide between the important possibilities that, (a) 95% of the cells have been killed and stopped metabolizing and the remainder are normal, or (b) all of the cells have been reduced to 5% of their former metabolism, and virtually all of them are dead, or (c) all the cells have been reduced in metabolic activity, but only a fraction, say, one-third, or perhaps none, of them have been killed For any study of mechanism, then, one of the first prerequisites is a sensitive method of estimating the proportion of bacteria killed. Such methods have not been used in any study of mode of action of surface active agents so far reported.

Obviously, the relation between metabolic inhibition and killing effect has not been adequately investigated. Inactivation of metabolic systems probably has some connection with bacterial death, although one cannot judge whether it is the actual cause of eventual death, or is merely one of the results of death brought about by another type of injury. Certainly, each view appears in itself plausible enough to be considered

MECHANISM OF THE METABOLIC INHIBITION

Surface active agents are known to concentrate as oriented layers upon interfaces, and, in many cases, to denature or disarrange various proteins, including some of the enzymes. These properties are reviewed and discussed thoroughly in the other papers of this symposium. These known potentialities have served as the bases for what few hypotheses have been offered to explain the bactericidal action. Kuhn and Bielig made the definite proposal that the capacity to react with protein and dissociate protein conjugates brings about the inactivation of essential components of living bacteria and thus occasions the death

of the organisms.[3] Later, one of Kuhn's collaborators, Jerchel, observed an inhibition of glycolysis that occurred when surface active agents acted upon streptobacteria.[4] Meanwhile, Baker, Harrison, and Miller had described their experiments upon inhibition of bacterial metabolism[1] and then followed with articles describing the bactericidal action of several of the same agents.[2, 5] These authors proposed as a working hypothesis of bactericidal mechanism a two-fold action: first, a disorganization of the cell membrane by virtue of the surface activity of the agents; and, secondly, a denaturation of proteins essential to metabolism and growth. Neither of these actions was directly observable, although their data was not incompatible with the assumption of the second one.

From what we know of bacterial metabolism, it is safe to say that the disorganization of the cell membrane could, in itself, have a pronounced effect upon metabolism and growth. Furthermore, the concentrations of detergents capable of denaturing most ordinary soluble proteins are in a higher range than those necessary for killing of bacteria. Thus, Anson treated proteins with from one-fifth to more than equal their weight of detergent to produce denaturation.[6] Kuhn and Bielig[3] obtained borderline effects with about equal weights, increasing until the ratio of detergent to protein in some cases was 25 or 50. The latter authors used chiefly rather complex and rare conjugated proteins, with, in many cases, spectroscopic measurement of denaturation. Bacteria, with combining weights similar to those of pure proteins[7, 8] are killed by one-fiftieth to one-hundredth their weight of suitable cationic detergents and, at most, one-tenth to one-fifteenth their weight of anionic detergents. It can be estimated that the killing of *Staphylococcus* occurs with about one-twentieth as much detergent as the cells are able to adsorb, according to Valko and Dibblee (mentioned by Valko[9]). *Escherichia coli* requires more nearly a saturating level. One may conclude that the enzyme proteins or other vulnerable elements of the bacteria must have a higher sensitivity to surface active agents than do other proteins in general, including even the other proteins of the bacterial cell itself, with which they appear to compete successfully.

Since the breaking of any weak link in the metabolic chain may

[3] **Kuhn, R., & H. J. Bielig.** Ber deut. chem. Ges. **73**: 1080. 1940
[4] **Jerchel, D.** Ber. deut. chem. Ges. **75**: 75. 1942.
[5] **Baker, Z., R. W. Harrison & B. F. Miller.** J. Exp Med **74**: 621 1941
[6] **Anson, M. L.** J. Gen. Physiol. **23**: 239 1939.
[7] **McCalla, T. M.** J. Bact **41**: 775. 1941
[8] **Stearn, A. E., & E. W. Stearn.** J. Bact. **9**: 491. 1924.
[9] **Valko, E. I.** Ann. N. Y. Acad. Sci. **46**(6): 476. 1946.

cause a marked inhibition of total metabolism, let us suppose that such a critically sensitive enzyme exists and its activity is interfered with by small amounts of detergent. We would expect that increasing the amount of detergent would soon give an excess capable of altogether blocking the metabolic activity. TABLE 1 shows the percentage inhibi-

TABLE 1

INHIBITION OF BACTERIAL METABOLISM BY INCREASING CONCENTRATIONS OF SURFACE ACTIVE AGENTS

Agent	Ref.	Per cent inhibition of metabolism; Killing				
		µg./ml.	*Staph. aureus*	*Esch. coli*	*Proteus vulgaris*	*Lactobacillus*
Zephiran	5	17	70	80		79
		33	84 K	89		82, 92 K
		70	86	92		92, 91
		100	87 K	— K		92, 94
Zephiran	1	33	94 K	92	94	83 K
		333	92 K	93 K	86 K	93 K
Retarder LA	1	33	65, 95 K	95, 76	85, 87	87, 79 K
		333	91 K	93 K	83 K	89 K
Cetyl pyridinium chloride	1	33	93, 95	89, 65	83	81, 41
		333		91	92	84
Cetyl sulfate	1	33	4			85
		333	20			88
Tyrocidine hydrochloride (unpublished)		72	65 K90%	Tergitol-7 ref. 5	µg./ml.	
		108	89 K		33	90, 84* K
		162	94 K		67	96
		243	95 K		100	95 K
					333	87*, 92* K

K—complete killing observed at this concentration (lower concentrations incompletely or inadequately tested) For killing data of ref 1 see ref 2

* Reference 1

tion of metabolism in several concentration series. It will be seen that once the concentration of agent is high enough to produce an inhibition of 80–95% there is relatively little effect of further increases of concentration. Indeed, it is relatively uncommon to observe more than 95% inhibition. These data are not in good agreement with the hypothesis that there is an enzyme of key importance that is very sensitive to the surface active agents, although one may postulate that there is such a one and, in addition, a small residual metabolism operating via an alternative and subsidiary pathway.

Let us now consider what would be the effect if the agent disorganizes, not an enzyme protein, but rather some component or some arrangement of components in the bacterial cell membrane. If the membrane of a particular cell is sufficiently damaged, we should expect the intracellular constituents, such as enzymes, essential ions, coenzymes and intermediates to be lost into the suspending medium. It is known from experience with other cells that such an enormous dilution of the intracellular constituents will reduce the metabolic activity to a small fraction of its former value, since the full activity is dependent upon maintenance of these multifarious components within the very small volume of the intact cell. However, once the cells of a given suspension have been so damaged, no further dilution can occur when more agent is added, and further depression of metabolic activity will not be produced, at least until sufficient agent has been added to interfere in another way as perhaps by denaturing an enzyme. The latter type of effect might or might not supervene in a given case.

The data of TABLE 1 are accordingly in agreement with a hypothesis based upon damage to a cell membrane with, in general, little or no effect upon key enzymes. Most of the data of Baker, Harrison, and Miller,[1, 5] not given in the table, are in harmony with this picture, although, in a number of cases, the two concentrations tested do not allow us to demonstrate a plateau of inhibition, and there are, to be sure, a certain small proportion of cases in which 100% inhibition was reported. The important thing is that there are many cases, involving some of the most typical bactericidal agents, where a plateau is reached

EVIDENCE FOR CYTOLYTIC DAMAGE

Hemolytic and cytolytic effects of the natural surface active agents, bile salts, saponins and fatty acids, have been well studied. There are only a few reports of such effects of the synthetic detergents,[10,11, 12] but there is little doubt that they behave similarly. In hemolysis, the red blood cell is observed to lose its original form and to give up all its hemoglobin into the suspending medium. Heretofore, there appears to have been no attempt to discover whether an analogous process occurs with bacteria. What is ordinarily referred to as "bacteriolysis" is probably almost always an enzymatic process, frequently an autolysis, and only occurs in certain species, when death has been brought about under conditions favorable for enzyme action.

[10] Bayliss, M. J. Lab. and Clin. Med. 22: 700. 1937
[11] Schulman, J. H., & E. K. Rideal. Proc. Roy. Soc London B 122: 29 1937
[12] Höber, R., & J. Höber. J. Gen. Physiol. 25: 705. 1942

When bacteria are cytolyzed, we find seeping out into the surrounding medium, not hemoglobin, but the ions, coenzymes and intermediates mentioned above. These are not visible like hemoglobin, but chemical analyses have revealed them in the washings of various bacterial suspensions containing added agents. At the same time, dilutions of the

TABLE 2

EFFECT OF CATIONIC SURFACE ACTIVE AGENTS ON STAPHYLOCOCCI

Agent	Chemical Structure	Conc. µg./ml.	Killing effect	Per cent of P and N released	
				P	N
AM 1120	dodecylamine	500 200	++++ ++	105 55	100 20
AM 1160	hexadecylamine	500 200	+ –	5 5	0 0
Triton K-60	dodecyl dimethyl benzyl ammonium	310	–	10	5
Zephiran	alkyl dimethyl benzyl ammonium	165	+ + ++	80	70
Phemerol	p-tert-octyl phenoxy ethoxyethyl dimethyl benzyl ammonium	500	++++	115	90
Sapamine KW	quaternary ammonium fatty amide	860 500	+++ +	55 15	30 0
Ceepryn	hexadecyl pyridinium (pure)	500 125	++++ ++++	150 100	110 75
Fixanol	hexadecyl pyridinium (technical)	830 200	++++ ++	95 55	115 45
Tyrocidine	basic lipoid-soluble polypeptide	165	++++	105	130
Tyrocidine acetyl deriv.	neutral lipoid-soluble polypeptide	500 200	– –	40 10	30 5

suspensions were made and incubated in normal growth media to see whether killing had occurred. The extent of killing was estimated with moderate accuracy by noting the time and rate at which growth occurred. This made possible the comparison of killing with bacterial cytolysis represented in TABLES 2 to 4

From these tables, the conclusion is drawn that, whenever the surface active agent and its concentration are adequate to kill the bacteria, there is observed a leakage of nitrogen and phosphorus compounds out

TABLE 3
EFFECT OF ANIONIC SURFACE ACTIVE AGENTS ON STAPHYLOCOCCI

Agent	Chemical Structure	Conc. µg./ml.	Killing effect	Per cent of P and N released	
				P	N
Duponol C	lauryl sulfate	830 400	+++ +++	100 130	65
Santomerse	dodecyl benzenesulfonate	500	++++	140	90
Triton W-30	alkyl phenoxyethyl sulfate	485.	++++	110	105
Ultrawet A	polyalkyl benzenesulfonate	500	−	1	15
	naphthalene-β-sulfonate	830	−	0	5
Aerosol OT	dioctyl sulfosuccinate	500	++++	145	105
Aerosol 18	N-octadecyl sulfosuccinamate	500	+++	85	70
Aerosol 22	N-octadecyl-N-1,2-dicarboxy-ethyl-succinamate	500	++	80	80
Aerosol 25	diamyl-N-2-hydroxyethyl aspartate	500	+	0	20
	hexadecyl malonate	165	++++	100	70
	oleate	4150	++++	115	105
	palmitate	2500	+++	145	
	phenol	15000 4000 1000	++++ − −	105 25 5	250 230 5
Tricresol	o,m,p-cresol	5000 1000 200	++++ − −	90 5 5	250 75 10
	o-chlorophenol	5000 1000 200	++++ − −	100 10 5	90 0 0
ST-37	hexyl resorcinol	830 500	++++ ++++	95 105	125 110
Taurocholate	bile salt	520	−	5	30

of the cells. Some of the compounds tested need some explanation. Among the cationic substances is tyrocidine, a natural basic polypeptide which, by virtue of a large number of organic side chains (particularly

phenylalanyl and leucyl residues), has the hydrophobic-hydrophilic nature and effects upon surface tension of an ordinary detergent. The surface tension of a 0.01% aqueous solution is about 40 dynes/cm.[13] When this bactericidal (and hemolytic) substance is acetylated, the virtually neutral acetyl derivative no longer has the marked surface activity and its bactericidal activity is largely destroyed. Also, among the anionic substances are some simple phenols, not usually thought of as surface active agents. Perhaps because their tendency to be ori-

TABLE 4

EFFECT OF NON-IONIZED SURFACE ACTIVE AGENTS ON STAPHYLOCOCCI

Agent	Chemical Structure	Conc. µg./ml.	Killing effect	Per cent of P and N released	
				P	N
Triton NE	aryl alkyl polyether alcohol	520	—	15	4
S 325	polyglyceryl stearate	500		0	60
S 255	pentaerythritol stearate	500	—	0	20
S 307	nonaethylene glycol laurate	500		5	75
Span 20	sorbitan monolaurate	500*	—	0	125
Span 80	sorbitan monooleate	500*	—	0	70
Tween 20	polyoxyalkylene sorbitan laurate	500		5	50
Tween 80	polyoxyalkylene sorbitan oleate	500	—	0	65
Arlacel C	sorbitan monooleate (purified)	500*		0	75
Saponin	sterol glucoside	500	-	0	5

* Water dispersion of insoluble agent.

ented at surfaces is slight compared to that of the highly developed synthetic agents, the simpler phenolic antiseptics are effective on bacteria only at relatively high concentrations. It is probable that many more, and perhaps all, of the synthetic agents would have caused cytolytic injury and killed bacteria, if they had been given the opportunity to act in as high concentrations as the phenols. In any case, the data of TABLES 2 to 4 are not, in any sense, to be construed as comparisons of the efficacy of different preparations, but are merely correlations of killing with cytolytic effects under certain specific experimental conditions and at arbitrary total concentrations.

Cytolytic injury, therefore, does occur when surface active agents act upon staphylococci. Similar observations have been made with streptococci, *Escherichia coli,* and baker's yeast cells. On the other hand, this effect was not seen when the usual antiseptics, including hydrogen

[13] Heilman, D., & W. E. Herrell. Proc. Soc. Exp. Biol. Med. 47: 480. 1941.

peroxide, potassium permanganate, sodium perborate, formaldehyde, quinone, mercuric chloride, organic mercurials, benzoate, salicylate, acridine dyes, gentian violet, iodine, active chlorine compounds, boric acid, fluoride, and others, in amounts adequate to kill or prevent growth completely, acted upon staphylococci. As suggested before, such an injury is at least sufficient to explain the loss of metabolic activity and death caused by the surface active agents. The fact that a small residual metabolic activity usually remains, as already described, may be one indication that the complex enzyme systems are more commonly inactivated by dilution rather than denaturation. Another such indication is that fermentation by yeast extract, in which the active components are already "diluted" away from the cell, is not affected by those low concentrations of detergents that reduce the fermentation by yeast cells to a small fraction of the normal rate. If, on the other hand, one would maintain that the "residual" metabolism is due to the persistence of metabolic function in a small proportion of resistant cells, we can, at any rate, see from the completeness of killing recorded in TABLE 1 that these resistant cells must nevertheless have been killed. So here, again, we are unable to conclude that repression of metabolic systems has been the cause of death.

STAGES OF THE CYTOLYTIC INJURY

The general circumstances of bacterial killing by surface active agents suggest that combination of hydrophobic-hydrophilic ions with oppositely charged groups in the bacterium first occurs The well-documented effect of pH alteration, already mentioned, appears always to augment bactericidal action when it operates in the direction of favoring the production of groups having the appropriate charge in the cell constituents. In most cases, pH changes in the usual range could hardly be expected to produce appreciable changes in the dissociation of the highly ionized organic sulfates and quaternary ammonium compounds, although this had been suggested.[1] Valko and DuBois[14] have considered this phase of the interaction of cationic agents and bacteria in some detail. They presented evidence that, for a brief period of about five minutes after treatment with a cationic agent, the killing of the bacteria was "reversed" by lauryl sulfate, an anionic agent. Their interpretation of this effect is that the oppositely charged agents can interact and accordingly reverse the adsorption of cationic surface active agents upon negatively charged groups within

[14] Valko, E. I., & A. S. DuBois. J. Bact. 47: 15. 1944.

the bacterial cell—adsorption which would otherwise rapidly have caused death of the cells. Their data also suggest that two cations of different toxicity compete with each other in combining with the bacteria.

Unfortunately, here again, killing was judged by a method revealing only a 100% end point, so that the number of cells revived may have been negligibly small. Since, a few minutes later, no viable cells could be recovered under any circumstances, it can even be considered probable that very few remained alive even at the end of five minutes. This being so, it might be that, at five minutes, the "reversing" agent was merely added just in time to prevent the last of the cells from being killed, and actual reversal was not accomplished. When the agent was not added, these last few cells might have died in the subculture during the next few minutes.

The investigations of Baker, Harrison, and Miller[5] have indicated similar antagonistic effects, although even the most potent preventing agents, phospholipids or a non-ionizing detergent, were not observed to interfere with the action of the surface active agents unless added before them. In a later note, Miller, Abrams, Dorfman, and Klein[15] reported that protamines increased the susceptibility of Gram-negative bacteria to anionic detergents, possibly by its interaction with a natural inhibitory substance, similar to the phospholipids.

The combination of oppositely charged surface active ions and bacteria was also assumed by Albert[16] along lines foreshadowed by the work of Stearn and Stearn[8] and McCalla[7] on the combination of other antiseptic ions with bacteria. Some combination of this sort must, almost inevitably, precede a direct action of the agent upon the cell It must not be forgotten, however, that the chemical and physical nature of the hydrophobic portion of the agent will play a large part in determining whether the combination will be momentary or permanent and irreversible. If the water insolubility of this portion is the driving force that tends to crowd even minute amounts of the agent out to the boundaries of the aqueous phase, it is the specific affinities of the water insoluble groups for elements in the bacterial structure that enable it to become anchored at particular points in or upon the cell. By virtue of these properties, the surface active ion may have, in very low concentrations, effects upon bacteria which cannot be duplicated by simpler ions or even by closely related ions with, say, the same structure and a somewhat shorter alkyl group.

These considerations do not go very far in explaining differences be-

---

[15] **Miller, B. F., R. Abrams, A. Dorfman & M. Klein.** Science **96:** 428. 1942.
[16] **Albert, A.** Lancet **1942** (II). 633

tween different types of cell, nor do they give much indication of what part or constituent of the cell is affected. Kuhn and Jerchel[17] observed that surface active tetrazolium salts are reduced to red formazans by yeast and sedimented with the organisms in the centrifuge, describing the process as penetration into the interior of the living cells. It does not appear from the abstract that there has been any demonstration that the reduced compound is not merely adsorbed upon the cell surface. Certainly, the tetrazolium salt might have been reduced without penetration. In the classical Thunberg metabolism procedure, living bacteria and other cells rapidly reduce methylene blue and other dyes without being stained by them. In the case of the usual synthetic surface active agents, highly dissociated salts with varying chemical natures, the experience of cell physiology would tend to predict that they could not penetrate the cell as long as it is alive. It is reasonable, however, that when accumulated or adsorbed upon it in sufficient amount, they may be able to initiate damage to the membrane, whereupon, as already described, cell solutes will leak out. At this stage, the extraneous agents may naturally penetrate more completely and may or may not bring about injury to other cellular constituents.

Whether or not some metabolic enzymes are inactivated by the surface active agent, the membrane injury is the signal for the beginning of a series of enzymatic processes which may lead to the virtual dissolution of the cell. Since these processes are autolytic, it should not be said that the agent has "lysed" the bacteria, but rather that it has "initiated lysis." There are great differences in the rate and extent of autolysis in different bacterial species. As shown in TABLE 5, if staphylococci are allowed to stand with the poisons, there is a steady release of nitrogen and phosphorus and a gradual partial clearing ("lysis") of the suspension. If the same mixture is kept cold or treated with mercuric chloride, these enzymatic changes are minimized, the cell does not change in appearance, and indeed it would be found to stain normally. The initial amount of nitrogen and phosphorus released is always close to the amount extractable by trichloracetic acid, the standard extractant for cell solutes. In TABLES 2 to 4 we have many examples of more than 100% release of solutes, since the deliberately practical exposure, 20 minutes at 25° C., was sufficient to allow autolysis. The large amount of nitrogen compounds extracted by the simpler phenols may, perhaps, be attributed to the solubility of some proteins in these rather high concentrations of phenol rather than to extensive autolysis.

---

[17] Kuhn, R., & D. Jerchel. Ber. deut. chem. Ges. 74: 949. 1941; Chem. Abstr. 35: 6957. 1941.

TABLE 5
DEGRADATION OF STAPHYLOCOCCI INITIATED BY SURFACE ACTIVE AGENTS

Agent	Temp. °C	Time min.	Per cent of initially extractable N and P released from cells: Inorganic phosphorus	Total phosphorus	Total nitrogen	Lysis: Per cent initial turbidity
Tyrocidine 165 µg./ml.	0	30	71	72	68	
	25	5	99	94	106	
	25	30	111	127	139	
	25	90	131	254	254	
Tyrocidine 150 µg./ml.	0	5	107	99		100
	0	30	121	114		103
	0	150	126	117		106
	38	5	167	136		100
	38	30	207	204		98
	38	150	232	296		76
Ceepryn 150 µg./ml.	0	5	125	101		100
	0	30	128	100		105
	0	165	136	103		107
	38	5	170	121		100
	38	30	189	137		94
	38	150	238	189		74
Ceepryn 150 µg./ml. + $HgCl_2$ 1: 18,000	38	5	162	113		100
	38	150	186	125		105
Duponol C 500 µg./ml.	0	5	105	109		100
	0	30	123	120		101
	0	150	130	125		104
	38	5	144	145		100
	38	30	158	151		101
	38	150	168	174		100

CONCLUSIONS

1. The first stage of the interaction of surface active agents and bacteria may be pictured as a combination of surface active ions with oppositely charged sites upon the bacterial surface. This process may be prevented or perhaps even reversed through the competition of suitably constituted ions, such as phosphatides, other detergents, and also hydrogen and hydroxyl ions.

2. If the hydrophobic groups of the surface active agent have the appropriate affinity for the bacterial surface, adsorption of a small fraction of the maximum amount that can combine will result in irreversible damage to the cellular membrane, so that the total content of soluble

nitrogen and phosphorus compounds is released from the cell. This process appears to be the analogue for bacteria of the hemolysis of red blood cells by surface active agents. At this stage, the cells are dead and their metabolic activity is very low, although morphologically they appear unchanged. Maintenance of a low temperature or additions of certain enzyme poisons can keep them in this state for a limited time.

3. The cells, after this cytolytic injury, are no longer able to repair themselves and begin to autolyze, so that cell constituents break down enzymatically and nitrogen and phosphorus compounds are liberated in greatly increased quantities. The rate and extent of this autolysis are characteristic for each bacterial species and strain. It is possible that, at this stage, with at least some species and some surface active agents, sufficient of the agent will penetrate to denature the metabolic enzyme systems.

4. With certain species of bacteria, autolysis involves, in its secondary stages, destruction of the cellular structure and clearing or partial clearing of the suspension.

5. Low concentrations of surface active agents appear to kill bacteria only when they simultaneously initiate the changes outlined above. This cytolytic type of injury is not a noticeably important feature of the killing of bacteria by other types of antiseptics.

DISCUSSION OF THE PAPERS BY DOCTORS VALKO AND HOTCHKISS

Dr. A. W. Ralston (*Armour Laboratories, Chicago, Ill.*):

In comparing bactericidal activities of surface active compounds, Dr. Valko drew certain relationships between the ionic charges upon the surface active ions and the bactericidal effect, the two being related by means of a series of mathematical equations appearing in Dr. Valko's paper. The subsequent paper by Dr. Hotchkiss suggested that electrical forces are involved in the attraction of the surface active ion for the bacteria and postulated an actual attachment of the ion to the cell membrane which process occurs prior to an actual rupture of the membrane. The validity of these assumptions was supported by convincing and fundamental experimental work and is unquestioned by the writer. It was, however, pointed out that, in view of the tendency of surface active ions to associate and form ionic micelles, the possibility of the ionic micelles and not the simple ions being involved in such a process should be considered. This is particularly true in view of the fact that the ionic micelle is the highest charged particle present in aqueous systems of colloidal electrolytes, and it appears that considerations relating electrical effects to specific actions upon bacteria should not disregard the presence of these highly charged particles. This does not change the fundamental concept as presented by Doctors Valko and Hotchkiss, but will somewhat complicate its mathematical treatment.

In discussing this question, Dr. Valko pointed out that bactericidal activity is evidenced at much higher dilutions than the critical concentrations for micelle formation, and cited as an illustration the curves presented for the equivalent conductivities and transference numbers of the amine hydrochlorides. While this is

admitted, it should be borne in mind that micelle formation is evidenced in extremely dilute solutions and that the critical concentration is probably attributable to a solubility effect and not to a spontaneous formation of ionic micelles. The fact that bactericidal activity is apparent below the critical concentration, therefore, does not appear to be significant, since ionic micelles are also present below this concentration. In view of the above considerations, it is felt that attempts to correlate bactericidal activity with ionic charges should not ignore the presence of the highly charged ionic micelles as an active component of the system.

Dr. Orville Wyss (*Wallace and Tiernan Products, Inc., Belleville, N. J.*):

The results of studies on the mechanism of the antibacterial action of surface-active agents should be useful to those investigating the preparation of more powerful agents of this type. It appears that the antagonistic action is a summation to two characteristics of these compounds: (1) their tendency to concentrate in or about the bacterial cell, thus giving a higher concentration than is present in the menstruum, and (2) the absolute toxicity of the substance at the site of its action. It seems likely that it should be possible to simplify the study by separating these two effects so that, in any series of compounds, one could determine the maximum theoretical effectiveness. For example, progressive modification of a type of molecular structure resulted in an increase in the absolute toxicity of the substance, which was not evident from a study of the total antibacterial effect, because it was masked by a simultaneous decrease in the tendency of the modified molecules to concentrate on or in the bacterial cells. By studying simplified systems and with our increasing knowledge of the mechanism, it may be possible to hit upon the most effective compound of a series with a considerable saving in synthetic work

[See, also, comments by Dr. Harry Sobotka, page 508.]

SURFACE ACTIVE COMPOUNDS IN FLOTATION ORE DRESSING

By M. H. Hassialis

From School of Mines, Columbia University, New York, N. Y.

Excepting food, clothing and some building materials, the basic materials of present-day civilization are derived from the mineral matter comprising the earth's crust. These mineral crudes rarely occur in useful form, but are found chemically and/or mechanically admixed with other usually useless substances. Reduction of the ore to useful form is achieved by application thereto of one or more separation processes, of which the end product is a primary-consumer derivative. In general, the overall separation may be divided into two stages: The first attempts removal of the adverse effects of mechanical admixture by segregating the valuable mineral into a concentrate containing the bulk of it in relatively pure form; the second stage of separation, usually applied to a concentrate, yields the primary-consumer derivative by severing the chemical bonds which hold the desired constituent in undesired association. In engineering parlance, the first stage is known as milling, the second as smelting.

Typical is the case of lead. It is derived, principally, from the mineral, galena, a chemical compound with sulfur. Galena is found at scattered localities in the earth's crust admixed with one or more worthless (gangue) minerals and sometimes with other valuable minerals. The principal primary-consumer derivatives of lead are: the metal, for use in pipes, alloys, storage batteries, cable coverings, etc.; the peroxide, for use in storage batteries; and the basic carbonate (white lead), for use in paints. In order to reduce galena to one of these forms, say the metal, it is necessary to separate the lead from the chemically associated sulfur. One method employed is to mix the lead-bearing minerals with a charge of such composition that, when the mixture is melted, the worthless materials combine together to form a complex silicate (slag) while metallic lead is freed. The molten lead, being insoluble in and heavier than the molten slag, settles to form a second liquid phase which is then drawn from the reaction crucible and cast as crude lead. Direct application of this chemical-separation process to the lead ore is economically possible only when the galena occurs as large masses which can be mined in a state of relatively high purity. This is rarely the case. Galena generally occurs as small grains inti-

mately admixed with large quantities of gangue, such as quartz and limestone. Direct chemical separation of the lead in a low-grade ore is usually uneconomic, because of the higher freight and smelting costs resulting from the handling of a larger volume of material for each unit of lead extracted. In such cases, it is necessary to first subject the ore to a low-cost separation which segregates the galena grains from the admixed worthless minerals into a concentrate of small bulk, and then chemically separate the lead out of this concentrate.

The discovery and perfection of low-cost, high-recovery separation processes, whereby the value-bearing minerals are segregated into a concentrate of comparitively small bulk, is the general problem of ore-dressing. Ore dressing methods are essentially mechanical in nature and are based upon differences in the physico-chemical properties of mineral and gangue particles. Owing to these differences, the separation processes cause the mineral and gangue particles to follow different paths, e.g., galena may be separated from quartz grains by placing the mixture in a liquid of intermediate density; whereupon the denser galena particles sink while the less dense quartz particles float. Other properties which have been utilized are: color, luster, cleavage, magnetic permeability, interfacial resistance and interfacial chemical properties.

FLOTATION

Flotation is essentially a mechanical method of separation which utilizes the differences in the interfacial physico-chemical properties of mineral and gangue particles. Owing to such differences, it is possible, by interaction with organic reagents, to render water-repellent the surface of particles to-be-floated, while simultaneously maintaining hydrophilic the surface of other particles. By causing the hydrophobic particles to become more or less tenaciously attached to air bubbles, their effective specific gravity is decreased, whereupon they float in the separating medium, while the hydrophilic particles sink. To illustrate,—consider the application of flotation to a simple lead ore, consisting essentially of galena finely disseminated in dolomite. The ore also contains some pyrite and a trace of sphalerite. The ore is crushed and ground to pass a 48-mesh screen, at which size it is substantially completely liberated, i.e., individual grains are either galena or gangue. The comminuted ore is pulped with water sufficient to give a suspension containing 15% to 35% solids and then added to a flotation machine which, in its simplest form, consists of an open-topped box provided

with a rotary agitator. The stirrer is started and the following reagents introduced in order: sodium carbonate, 3–4 lb. per ton of solids; potassium ethyl xanthate, 0.2 lb. per ton; cresol, 0.1 lb. per ton. Soon after the addition of cresol, a galena-laden froth collects at the top of the pulp. This is skimmed until the froth becomes barren. Examination of the concentrate (float material) shows it to contain substantially all of the galena and only small amounts of gangue, while the tailing (the non-floated material) contains substantially all of the gangue and only a trace of galena. The same results may be obtained by using different reagents and/or a different machine. Also, by a suitable choice of reagents, it is possible to make the gangue report in the froth while the galena remains behind.

ELEMENTS OF THE FLOTATION PROCESS

The essential elements of the flotation process are: (1) collection, (2) conditioning, (3) frothing, (4) bubble-attachment and (5) levitation. Collection is a selective change in the surface properties of the particles as a result of which the particles to-be-floated are made water-repellent while other particles remain water avid. Reagents effective in producing such a selective change in surface properties are termed collectors. Conditioning consists of such operations as control the state of the particle surfaces and the state of the aqeous solution. In a more formal development of the subject, collection would be considered as a subdivision of conditioning. Conditioning which aids collection is termed activation, while conditioning which prevents collection is called depression. Frothing is the production and maintenance of a collection of bubbles at the upper surface of a liquid, effective reagents being termed frothers. Bubble attachment refers to the process whereby air bubbles become attached to water-repellant particles Levitation is the act of buoying to the pulp surface, by means of air bubbles, particles attached thereto.

COLLECTION

The sole object of collection is to effect selective water repellence of the particles to-be-floated, in order that air bubbles may attach thereto. Practically all naturally occurring minerals have hydrophilic surfaces, except only such minerals as gilsonite and ozocerite, which are essentially hydrocarbons with a high carbon to hydrogen ratio.[1] However,

[1] Taggart, A. F., G. R. M. del Giudice, A. Sadler & M. D. Hassialis. Trans. Am. Inst. Mining and Met. Engs. 194: 180. 1943

the normal hydrophilic character of mineral surfaces may be altered by treatment with a suitable collector. The alteration in surface properties is most readily followed in the bubble machine, developed and perfected by Taggart and his co-workers.[2] A clean, polished mineral specimen is immersed in a solution of the collector, contained in a rectangular all-glass cell, the ensemble being placed on the stage of an optical bench. An air bubble, held captive in a cup hollowed in the lower immersed end of a vertical glass rod, is brought into pressure contact with the mineral surface and there maintained for about a minute. At the end of the pressure-contact time, the bubble is slowly removed, and its performance followed by observing its image in the ground glass of the camera. The bubble peels cleanly off hydrophilic surfaces, but is more or less tenaciously held by hydrophobic surfaces. In the latter case, a definite contact angle is developed.

Wark and Cox[3] showed that the contact angle developed with a particular collector is independent of the nature of the mineral, from which it may be argued that the surface is identically the same in all cases. Taggart, Taylor and Knoll[4] found that, in all cases tested, alteration of the surface properties is accompanied by removal of collector from solution, whence it follows that collector probably concentrates at the mineral surface. This was checked by dissolving off the mineral surface a crystallizable compound which analyzed as a metallic salt of the collector. It was further postulated that the collector is oriented at the mineral surface with the hydrocarbon part of the molecule away from the mineral. This postulate is supported by the fact that all collectors with the same normal chain length give the same contact angle,[3] e.g., ethyl xanthate, ethyl mercaptan, diethyl thiophosphoric acid and diethyl dithio carbamate. Since behavior is the same, regardless of mineral and of the chemical radical wherein substitution of the hydrocarbon radical takes place, and since water repellency does not develop in the absence of the hydrocarbon group, it is strongly implied that the surface of water repellency is actually the hydrocarbon-water interface. This postulate is supported additionally by the fact that a solid monomolecular film of stearic acid is water repellent on the hydrocarbon side and water avid on the carboxyl side.[5] Finally, if the orientation postulated does exist, it should be possible

[2] Taggart, A. F., T. C. T[illegible] Trans. Am. Inst. Min. and Met. Engrs. 87: 285. 1930; C. H. M. del [illegible] Eng. and Min. J. 137: 291. 1936.
[3] Wark, I. W., & K. B. Cox. Trans. Am. Inst. Min. & Met. Engrs. 112: 189.
[4] Taggart, A. F., T. C. Taylor & A. Knoll. Ibid. 87: 217. 1930.
[5] Langmuir, I. Trans. Far. Soc. 15: 69. 1920. Blodgett, K. J. Am. Chem. Soc. 56: 495. 1934; Ibid. 57: 1007. 1935.

to destroy the water repellency of an oriented film by locating a solubilizing group, such as a hydroxyl group, in the extreme end of the hydrocarbon part of the molecule. Thus, the water repellency imparted to the surface of galena by diphenyl thiourea should be missing when p-dihydroxy diphenyl thiourea is used, providing the latter compound also concentrates at the mineral surface. This is the case. A similar result was obtained by Taylor and Knoll[6] when glycol xanthate was used in place of ethyl xanthate.

The mechanism of the concentration of collector at the mineral surface has been the subject of much discussion in the literature of flotation. One school of thought postulates adsorption of the collector molecule at the mineral surface. The difficulty in this position lies in the fact that collectors are usually highly ionized, and in the failure to obtain adsorption isotherms. The final blow to this position is to be found in the exact stoichiometry obeyed by the collection of galena by ethyl xanthate. Taggart, Taylor and Knoll[4] showed that, when galena powder is suspended in a solution of potassium ethyl xanthate, shaken, filtered and the filtrate tested for EtX^-, there is a marked decrease in xanthate ion concentration, while there is no decrease in K^+ concentration. The filtrate is also found to contain $SO_4^=$, $CO_3^=$ and $S_nO_m^=$ ions, which are stoichiometrically equivalent to the lost EtX^-. The presence of $SO_4^=$, $CO_3^=$ and $S_nO_m^=$ is explicable on the basis of rapid surface oxidation of galena by the atmosphere. Thus, when the above cited abstraction test is repeated in a non-oxidizing atmosphere with galena powder whose surface has been thoroughly cleansed of all oxidation products, no abstraction of EtX^- is noted and no $SO_4^=$ $CO_3^=$, and/or $S_nO_m^=$ appear in the filtrate. This and similar tests led these authors to propose a chemical theory of flotation.

The theory of Taggart, Taylor and Knoll postulates chemical reaction of the collector with the mineral surface, resulting in the formation of a more insoluble compound than any existing thereat prior to such interaction and orientation of the compound formed, in consequence of which a hydrocarbon-like surface is presented to the liquid phase. Thus, in the tests cited above, it is postulated that EtX^- ions from solution react with lead ions in the lattice of galena, the resulting $Pb(EtX)_2$ being so oriented as to present the ethyl groups to the liquid phase. A recent objection to this mechanism was raised by the fact that $Pb(EtX)_2$ is a collector for galena from which it is argued that the true mechanism is reaction in solution between Pb^{++} and EtX^-

[6] Taylor, T. C., & A. Knoll. Trans. Am. Inst. Min. & Met. Engs: 112: 394. 1934.

to form $Pb(EtX)_2$, which is then adsorbed onto the surface of galena. This criticism is refuted by the fact that galena abstracts EtX^- from a solution of radioactive $Pb(EtX)_2$ but does not remove any of the radioactivity which is due to the lead.[7]

If the chemical theory of flotation be accepted, it follows that a collector must be capable of ionization and able to form a relatively insoluble salt with the anion or cation of the mineral surface and that, when oriented at the surface, it presents a hydrocarbon radical containing a number of carbon atoms sufficient to impart the requisite water repellency. Most known collectors are of this type. They possess a polar group which has an ionizable atom or radical and a non-polar, hydrocarbon-like end. The sole exceptions of practical importance are neutral oils, which are excellent collectors for a limited class of minerals, i e., sulphur, graphite, coal, molybdenite and orpiment. In this case, there is evidence of mutual solubility of oil and mineral, resulting in the attachment of a thin film of oil at the surface. It is not known whether this film is oriented.

Collectors are classified as anionic or cationic according as the hydrocarbon group is in the anion or cation of the collector. All collectors, excepting the neutral oils, are surface-active compounds, for they are polar compounds which function by concentration and orientation at the mineral-liquid interface. The following are the most commonly used anionic collectors: The xanthates, $RO{-}\overset{\overset{\displaystyle S}{\|}}{C}{-}SM$, where $M = Na$ or K and $R =$ alkyl radical, containing two to six carbon atoms; aryl and higher alkyl radicals have been used (Effectiveness of these compounds, as collectors, increases with the number of carbon atoms in the alkyl group); the thiophosphates, $\begin{matrix} RO \\ RO \end{matrix}\!\!>\!P\!\!<\!\!\begin{matrix} {\scriptstyle //}S \\ SM \end{matrix}$, where $M = Na$, K or NH_4 and R is an alkyl or aryl group, the most common being ethyl, isopropyl, sec-butyl, amyl and methylphenyl; the mercaptans RSH, where R is an alkyl or aryl group containing two or more carbon atoms; the thioureas, $(RNH)_2C = S$, where R is most commonly phenyl (This compound ionizes in the enol form); the carboxylic acids, RCOOH, of which the most commonly used are oleic and palmitic. The crude oils used in practice contain many other carboxylic acids. The cationic collectors have only recently been used in flotation and principally for

[7] Unpublished experimental work by the author.

the flotation of silica, silicates and scheelite, although laboratory tests indicate a much broader field of applicability. The most commonly used amine is the octadecyl, others are primary, secondary, or tertiary amines with about eight or more carbon atoms, quaternary ammonium salts and the like. There is some evidence to indicate that the amine may be concentrated at the mineral-liquid interface by complex formation as well as salt formation with the anion of the mineral.

CONDITIONING

The surface activity of a compound, in the flotation sense, i.e., its collector properties, depends not only upon the chemical structure of the compound, but also upon the nature of the mineral and of the liquid in contact therewith. Thus, potassium ethyl xanthate is a collector for galena, when the liquid phase is a water solution of xanthate. However, when the liquid phase also contains sodium hydroxide in an amount sufficient to raise the pH to about 10.5, xanthate no longer collects galena. Similarly, xanthate does not concentrate and orient at the surface of galena when the liquid phase contains K_2CrO_4, dihydroxydiphenyl thiourea or H^+ sufficient to give a pH of about 4.0. Such reagents, which function to prevent collection, are termed depressants. On the other hand, activators function to promote collection, e.g., sphalerite, which is not collected by ethyl xanthate, may be activated by preconditioning with Cu^{++}. Barite is not collected by lauryl amine at low concentrations, but may be activated by the addition of $SO_4^{=}$, etc.

The mechanics of conditioning apears to be explicable in terms of solution chemistry, though often this chemistry becomes a Procrustean berth for the facts. The use of OH^- as a depressant for galena is thought to be due to the formation of plumbate on the surface, which effectively protects the surface lead ions from reaction with xanthate. The chief difficulty with this explanation is the high solubility of sodium plumbate. More reasonable is the explanation for the depressant action of dihydroxydiphenyl thiourea. In this case lead xanthate is more soluble than lead dihydroxydiphenyl thiourea and, since the latter is hydrophilic, depression occurs. The depressing effect of H^+ may be ascribed to the formation of the very insoluble and unstable xanthic acid. The activation of sphalerite by Cu^{++} is said to result from the exchange of Cu^{++} with Zn^{++} and the formation of an insoluble copper sulphide. The difficulty lies in the implicit assumption that is made, i.e., Cu^{++} penetrates a film of $ZnSO_4 + ZnS_nO_m$ to reach $S^{=}$. Subsequently, this thin film of CuS, which is probably monomolecular, must

oxidize before reaction with xanthate can take place. The mechanism of such colloid depressants as starch, glue, gelatine, tannin, saponin, etc., and such dyes as Congo red, methylene blue, Hoffman violet, indigotine, sulphonated ingrosine, alizarin saphirol B, etc., is even more difficult of explanation. In some instances, it would appear that definite compound formation takes place at the mineral surface, the oriented compound presenting a hydrophilic surface. In other cases, it is thought that the micelles precipitate on the mineral surface to form static neutralization complexes or floccules between charged ionized mineral surface and ionized groups at the surface of the micelle. It is also probable that many of these depressants have an effect through the bubble surface.

Many conditioners have both activator and depressor properties. This is not surprising, if we remember that behavior of the interface depends upon the mineral, the liquid phase and the reagent. Thus, sodium hexametaphosphate acts as a depressant for apatite with oleic acid,[8] but as an activator for hematite with the same acid.[9] In the former case, it is probable that the calcium complex of the hexametaphosphate is formed and held at the mineral surface, which is then protected from interaction with the oleate ions. In the latter case, it might be conjectured that some iron salt is formed which makes available a higher concentration of iron ions for interaction with the oleic acid. The dispersing action of hexametaphosphate and other depressants may also be used to depress an undesired mineral; thus, silicates and, to a lesser degree, heavy-metal sulphides and oxides are deflocculated by hexametaphosphate and, since a mineral to-be-floated is invariably flocculated, depression is effected.

The ultimate objective of mineral-surface preparation is the discovery of specific collectors for each mineral; failing this, the discovery of conditioners which make specific the action of the collector. This is a problem of increasing importance as the high-grade ore deposits become depleted, leaving as primary sources only the low-grade and/or complex ores. These ores usually contain two or more valuable minerals in relatively small percentages, the minerals being finely disseminated through the gangue. It is necessary to produce separate concentrates of each of the valuable minerals and, since the dissemination is fine, this invariably implies separation by flotation, i.e., differential flotation. The importance of specific collection becomes apparent.

[8] Rose, E. H., & W. T. MacDonald. U. S. Patent 2,040,187. 1936.
[9] Keck, W. E., G. G. Eggleston & W. W. Lowry. Am. Inst. Min. and Met. Engrs. Tech. Pub. 762. 1937.

FROTHING

A froth, in flotation, is a relatively long-lived collection of bubbles at and above the surface of a liquid. The longevity or persistence of a froth depends upon its ability to resist strains. Pure liquids do not froth; i.e., froths of pure liquids have negligible persistence. On the other hand, solutions of certain reagents known as frothers give froths with good persistence characteristics. Two physical properties are responsible for the ability of films of liquid to resist strains, viscosity and surface tension.

Addition of a solute to a pure liquid either raises or lowers its surface free energy. According to Gibbs' equation:

$$\frac{d\gamma}{dc} = -RT\frac{U}{c}$$

where γ = surface tension, c = concentration of solute, R = gas constant, T = absolute temperature and U = concentration of solute at the interface, when the surface tension decreases with the concentration of solute, the solute is adsorbed at the interface. Conversely, when the surface tension increases, the solute is negatively adsorbed; i.e., the concentration of solute at the surface is lower than its concentration in the bulk of the solution. Organic solutes in water generally decrease its surface tension. Conversely, inorganic solutes invariably raise the surface tension. To explain the ability of a film of solution to resist strain, consider the case of a positively adsorbed frother. When the filmed is strained, the surface of the film is increased by bringing liquid from the interior to the surface. Since the bulk liquid has a lower concentration of solute, the new surface has a higher surface tension; hence, it offers greater resistance to strain. After a short interval of time, solute molecules diffusing into the surface reestablish the relation of U to c and the surface tension falls. Hence, if the original strain was not sufficient to rupture the strengthened film and did not persist too long, the film remains unbroken. In the case of an inorganic solute, extension of the film concentrates solute in the surface, which causes the surface tension to increase and thus increases resistance to breakage. In either case, maximum resistance is obtained when the solute concentration corresponds to that portion of the surface tension-concentration curve which exhibits the maximum rate of change of surface tension with concentration.

The viscosity of a froth as a factor in its ability to resist strain is well recognized. However, the exact meaning of the term, viscosity, when applied to a film, is in a state of confusion. In some instances,

the rate of drainage of liquid from a froth purports to measure the viscosity thereof. In other instances, resistance offered to the passage of a body there-through is used as a measure of viscosity. The term, surface viscosity, has also been applied, and with more definite justification, to the resistance to flow of a two-dimensional liquid through the surface-slit viscometer of Bresler and Talmud.[10] Ill-defined though the term may be, it appears that differential movement of liquid within a film produced by a strain is resisted by a force which we shall term film viscosity. It appears that film viscosity depends upon the nature of the frother and age of the solution. A possible explanation may be found in the observation that the surface tension of the solution suffers negligible change with time, whereas its Tyndal cone increases in intensity.[7] If the frother decomposes or oxidizes upon standing to give solid reaction products which report in the air-liquid interface, the thickness of the film liquid is decreased and liquid draining out of such a film must follow in part, the labyrinthine path resulting from the presence of these products in the surface. Finally, as the film thickness is further decreased, solid friction between solid particles in the opposite surfaces of the film, comes into play. If decomposition of frother results in a relatively viscous oil, concentration of the oil at the surface increases film viscosity. In some cases, a neutral oil is compounded with the frother to give high film viscosity and, by controlling the amount of admixed oil, the operator may control viscosity. A froth laden with selected mineral particles which are located in the air-liquid interface exhibits high film viscosity. Such a froth may show extreme resistance to strain, thus, in one case, it is reported that a mineral-laden froth supported a shovel. The close packing of the mineral particles and complete armoring of the bubble is shown by the fact that the dried froth does not disintegrate but retains its structure.

The most commonly used frothers are surface-active compounds which concentrate and orient at the air-liquid interface, orientation being induced by a molecular structure composed of a polar solubilizing group and a hydrocarbon radical of at least five carbon atoms. Solubility of good frothers normally ranges from 0.001% to about 4%. It is the resultant of the opposing tendencies of the polar and non-polar parts of the molecule. The solubilizing effect of polar groups in alkyl-type frothers appears to be $COOH > NH_2 > OH > CO$; in the aryl-type, $OH > NH_2 > CO > COOH$. Insolubility is dependent primarily upon the number of carbon atoms in the non-polar group. Frothers used in

[10] Bresler, S. E., & D. L. Talmud. Physikal. Z. Sovietunion 4: 864. 1933.

flotation should have little or no collecting action, so that the operator may control frothing independently of collection. They should be active frothers at low concentrations and should not be affected by the presence of other solutes. They should yield froths persistent enough to permit separation, but not so persistent that subsequent handling of concentrates becomes a major mill problem. Finally, they should not adsorb soaps or other organic colloids to the exclusion of solids. The most common frothing agents used in mills are: pine oil, eucalyptus oil, liquid Aerofloats, cresylic acid soaps, mixtures of aliphatic alcohols and sulphates and sulphonates of long-chain aliphatic alcohols.

BUBBLE ATTACHMENT

Two types of mineral-bubble attachment have been recognized as important in flotation: the precipitated-bubble type, wherein the mineral is located in the air-liquid interface and the preformed-bubble type, where the mineral is completely in the liquid phase contiguous to the interface. The former type predominates in flotation machines where aeration of the pulp is obtained by agitation, the latter in the so-called pneumatic machines where agitation and aeration of the pulp is simultaneously achieved by bubbling air through the pulp.

PREFORMED-BUBBLE ATTACHMENT

Preformed-bubble attachment is obtained in a wide variety of machines. Essentially, they are all alike in the fact that air is introduced into the pulp at the bottom through a porous mat or similar device. As the air bubbles rise, they come in contact with the particles in the pulp. The hydrophilic particles, upon collision with a rising bubble, slide down the bubble wall until the equatorial plane is reached and then drop off. The hydrophobic particles, on the other hand, continue to slide on the bubble wall until the lower pole is reached and remain there until a number of particles are collected, when the entire collection falls. This difference in behavior means that hydrophobic particles spend a longer time with the bubbles and, since these are rising, a net rising velocity is imparted to the particles. By proper control of the gas volume through the cell, the hydrophilic particles have a net settling velocity. Hence, separation of hydrophobic from hydrophilic particles is achieved.

The exact nature of this attachment is not known. It is known experimentally that, in bombardment of a captive bubble by hydrophobic particles, the force with which the particle presses against the bubble

is insufficient to locate the particle in the air-liquid interface. However, if the bubble is pressed against hydrophobic particles and the contact maintained for a short interval of time, the particles go into the air-liquid interface. Taggart has conjectured that, in the preformed-bubble type of attachment, the mineral is located in a liquid-liquid interface formed between the main body of liquid and a second liquid phase consisting of adsorbed frother and oil. The oil is introduced into the pulp by the ore and consists principally of lubricating oils used to service mining and milling equipment. Characteristic of this type of attachment are: the ease with which attachment is broken; incomplete coverage of the froth with mineral; large bubble size of froth and lack of froth persistence.

PRECIPITATED-BUBBLE TYPE OF ATTACHMENT

This type of attachment is obtained by local supersaturation of the liquid with the gas, brought about by increased temperature, reduced pressure and chemical or electrochemical generation. It can be shown by thermodynamic reasoning that the system is at minimum free energy with the gas precipitated at a hydrophobic interface. The location of the mineral surface in the gas-liquid interface may be demonstrated by causing water vapor within the bubble to condense. It does so on the mineral surface, forming a droplet of water which exhibits a characteristic contact angle.[7] In the agitation-type flotation machines, aeration of the pulp is achieved by a high-speed stirrer. Owing to the high speed, a vortex is formed about the stirrer shaft; closure of liquid over the apex of the vortex entrains air which is subsequently broken up into minute bubbles. Liquid and air, at the front of the impeller blade, are subject to increased pressure which aids solution of air in liquid. Conversely, liquid at the rear of the blade is in a zone of reduced pressure, which causes a supersaturation of gas. This supersaturation is relieved by precipitation of the gas on the surface of any hydrophobic particles in the immediate vicinity. This type of attachment is characterized by tenacity of attachment, more or less complete coverage of bubble by mineral, small bubble size and great persistence of froth.

The substantially complete coverage of the bubbles in the precipitated-gas type of attachment leads to the viscosity effects aforementioned and to solid friction between armor coatings. In consequence, hydrophilic particles are entrained and it becomes necessary to allow froth collected at the surface of the pulp sufficient time to cleanse itself

of tailing particles. The cleansing action is aided by a lively froth, i.e., a froth whose upper surface is in active coalescence, for such coalescence produces showering of mineral and gangue load which is cleaned by underlying bubbles. Some mills have found it expedient to control froth characteristics by using a mixture of two frothers, one of which produces a brittle froth, while the other makes a tough froth, control being achieved by varying the proportion.

LEVITATION

The actual lifting of hydrophobic particles to the surface of the pulp is, of course, due to the reduced overall density of the bubble-mineral aggregate. So long as the overall density of the aggregate is less than that of the pulp, buoyancy is the driving force. The ability of a bubble to lift a hydrophobic particle depends upon the nature of the attachment, the size, weight and shape of the particle, and upon the contact angle.

Generally speaking, there is a maximum size of particle that can be lifted in either type of attachment, the size being smaller in the case of the preformed-bubble type. Other things being equal, the smaller the density of the particle, the larger the maximum size. The importance of shape may be judged by the following facts. A flat particle may be lifted by a bubble when the flat side is presented to the bubble. However, when a thin side is attached to the bubble, the particle usually breaks away. The effect of increasing contact angle is to permit lifting of larger particles, other variables being kept constant. The qualitative relationships aforesaid are difficult to quantify. Certainly, in the case of the preformed-bubble type of attachment, little is known. In the precipitated-bubble type, the problem is directly related to the classical problems of the floating needle and disc. It can be shown that the weight the disc can support is, in accordance with Archimedes' principle, equal to the weight of liquid displaced below the general liquid surface, and the extent of displacement is controlled by the surface tension of the liquid and the contact angle of the disc surface.

RESUMÉ

In synthesis, the coaction of the process elements discussed above constitutes the process of flotation. Flotation is a process for the separation of finely divided mineral mixtures. The separation is effected by so treating the pulp with reagents that the surface of particles-to-be-floated becomes hydrophobic while the surface of other particles

is hydrophilic; then, either by causing gas to precipitate on or come in contact with the hydrophobic-surfaced particles, a preferential selection of these particles is made. The resulting bubble-mineral aggregates are buoyed to the surface of the pulp, whence they are removed by scraping.

The role of surface-active agents in flotation is of paramount importance, for they are used in all capacities, i.e., as collectors, conditioners and frothers. It is true that some of the reagents mentioned are not normally considered as surface active, when judged by restricted meaning implied by their classification according to use, i.e., wetting, detergent, dispersing, emulsifying, etc. Even in this restricted sense, surface-active agents are all important. Their greatest role has yet to be played, i.e., in the field of differential flotation. The close control of collection and frothing required in the differential flotation of metallic and non-metallic ores demands better understanding of the mechanics of the process and of the control of the various stages by proper selection of reagents. This, in turn, requires the availability of a large number of surface-active agents and a knowledge of their various characteristics. The recent surface consciousness of the chemical industry bids well to supply this need.

DISCUSSION OF THE PAPERS

January 27, 1945

Dr. Harry Sobotka (*Mt. Sinai Hospital, New York, N. Y.*):

Today's discussions are mostly concerned with the effect of synthetic surface active compounds on cells, tissues, and organisms. It may be appropriate to mention that animals and plants themselves provide a number of surface active substances; and that, in fact, Nature has given us important hints in the form of such models as the bile acids and the saponins. These substances, developed in the course of evolution, have their assigned tasks in animal and plant physiology. One of the main purposes of bile is the emulsification of fat and lipoids in food prior to its absorption in the intestine. Bile supplements the action of lipolytic and perhaps also of some other enzymes by emulsification of the respective substrate. Beyond their physiological scope, bile acids exert what is known as cytotoxic effects, i.e., they are poisonous for cells. The mode of action of these effects is based on surface activity, accumulation of the poison in surfaces and interfaces of the cell, changing permeability of membranes and altogether upsetting the equilibria at such two-dimensional structures. This toxic action is, in turn, utilized, e.g., for the lysis of pneumococcus cells as a therapeutic or diagnostic measure. It is well to keep these effects in mind, in view of the possible application of synthetic surface active substances in cosmetic and pharmaceutical preparations.

Other surface active substances are found in plants, the cardiac glucosides and particularly the saponins, which, like the bile acids, are steroids or have a five-ring carbon skeleton closely related to the four-ring carbon skeleton of the steroids. The bitter taste common to all these substances may be due to a general reaction of the taste organ to surface active substances. Their various toxic effects are closely linked to their surface activity. The specificity of their action is governed

by details of their chemical constitution, which gives them affinity and directs them to certain tissues, cells, and membranes.

For most of these substances, the same architectural plan holds as for synthetic surface active compounds: they contain a hydrophobic portion and a hydrophilic portion or group. The hydrophobic group in these natural products is not an aliphatic chain, but the annellated hydroaromatic ring system of the steroids. The carboxyl group, as such, seems to be insufficient as hydrophilic pole, and we see the bile acids conjugated with the amino acids glycine or taurine to "float" them, i.e., to make them watersoluble. The alcohol, scymnol, in shark bile is conjugated with sulfuric acid. In a similar fashion, the aglucons of the saponins are conjugated with glucose and certain desoxysugars to render them water soluble. The related toad poisons are conjugated with suberic acid and arginine.

The process of conjugation is not confined to these products of the animal body itself, but is also called into action whenever the body wishes to get rid of nonphysiological intruders The conjugation of benzoic acid and its derivatives with glycine to hippuric acid, or of naphthalene as a cysteine derivative, are usually considered as means to make the toxic substance watersoluble, but the surface activity of the conjugated detoxification products should be studied, since it might play an important role in their renal excretion. The disposition of hydroxy groups on one front of the steroid skeleton in bile acids, and of methyl groups on the other front, leaves much room for speculation. The specific effects of these individual groups on the various surface effects still await an explanation.

This brings me to a rather skeptical remark that was made here yesterday regarding the present knowledge of the relation of chemical structure to surface activity. Although this correlation is still in an empirical state, one may already discern possibilities for systematic approach. The technically interesting characteristics of the compounds under discussion, their detergent, emulsifying, wetting, and foaming characteristics are the product of several physical properties, amongst which surface tension is only one, in addition to their properties as a solute as well as a solvent · diffusion constants, electrolytic dissociation constants, coordinative affinity, wetting angles on contact with fluids, and geometric, i.e., steric characteristics of the molecule. Now, when we consider other physical properties of a molecule, say, its absorption spectrum, we do not hesitate to ascribe certain bands to definite parts or sectors of the molecule. Similar considerations are applied to electrical properties. I think Irving Langmuir was the first one, or one of the first, to consider surface tension, too, as what one might call a submolecular property, a property that may be ascribed to, or projected upon, a definite sector of a molecule, and you will perhaps remember the very fruitful analogy which he drew between the surface and interface tensions of a lens of paraffin oil with that of a mono-molecular fatty acid layer. I wonder whether our present knowledge of van der Waal's forces would not yet suffice to tackle, e.g., the problem of such maxima or minima of surface forces as we were shown yesterday to be observed around the C_{12} or C_{14}- members in certain homologous aliphatic series.

Finally, I should like to add another example to the subject of inflection points on surface tension/concentration curves. Ettisch, in some unpublished work, showed that glycocholic acid in distilled water or in mildly acid solution (up to N/1000 HCl) produces a distinct minimum of surface tension in N/500 — N/100 solution, but no minimum in alkaline solution of more than N/1000 NaOH; all curves then converge towards a value of 48 dynes/cm. for higher bile salt concentrations. I wonder whether this effect should be ascribed to a contamination as Dr. Shedlovsky suggested for similar cases.

THE INDUSTRIAL USE OF SURFACE ACTIVE AGENTS

By Robert R. Ackley

Mellon Institute of Industrial Research, Pittsburgh, Pennsylvania

An ultimate purpose of the study of the properties of any series of chemical compounds is the acquisition of information leading to the best possible utilization in industrial processes. Indeed, it is the impetus lent by commercial exploitation which is responsible for the rapid development in many fields. Although many of the published fundamental papers in the field of surface active agents are the product of research of academic institutions, it is safe to surmise that an even greater amount of investigation has been performed in industrially supported laboratories. Such studies are intended, on the one hand, to assist the manufacturer in the development and production of salable merchandise on which a return may be realized, and, on the other hand, to improve processes in which they are employed by the purchaser.

There are very few industries in which surface active agents are not employed and new uses are constantly being developed. Gradual improvements in the nature of the agents used and in the methods for their application are being made constantly. When a manufacturer makes available a new agent, it is quickly tested by many users for a very wide variety of applications. Similarly, as new applications become known, manufacturers attempt the development of agents with maximum suitability.

Investigators in the field are continuously working towards a better understanding of the relations between properties and constitution as well as the interpretation of efficiency in terms of the standard, readily measured, physical properties. Their researches are pointing towards the ideal situation wherein, for a given application, the important physical properties may be determined, and, from the known properties of the commercially available agents, the proper one selected.

If this highly desirable condition is to be realized, further investigation is necessary in four directions:

1. The improvement of methods for determining the physical properties of solutions of surface active agents.
2. The assignment of satisfactory mechanisms to account for the observed properties.

3. The correlation of properties and mechanism with molecular geometry.
4. The correlation of properties and mechanism with end-use.

Each of these phases has already been studied by many investigators. Even in the first problem, involving the methods for measurement of properties, there is still some disagreement as to the best methods. The second step, requiring interpretation, is in a state of considerable confusion, as in the explanation of the minimum in surface tension curves. As we have seen from the paper of Price,[1] only the broadest generalizations are possible for the relation of constitution to properties. Even less progress has been made in the understanding of suitability on the basis of the observed common physical properties.

It is with this last phase of the problem that this paper is concerned.

DIFFICULTIES IN THE PREDICTION OF SUITABILITY

In effect, we are interested in a knowledge of the significance of the properties of solutions of surface active agents in order that we may more readily predict their end-use suitability, or their relative efficiency for a given technical application.

The concept of efficiency for industrial use is different from that in a purely academic investigation. In the latter case, the units in which efficiency is measured are either those of mass, indicating, for example, the percentage concentration of a solution required to perform a given function, or units of molar concentration, which are most useful for interpretation of mechanism. Industrially, however, the units are either degree of excellence of effect obtained, relative cost, or ease of use. Perhaps the most important of these considerations is the economic efficiency, or the work done per dollar spent. There is probably as much investigation, today, leading towards performing an operation more cheaply as is directed towards improving the effect obtained.

In making a survey of the use of surface active agents in industry, it might, at first, appear that the relative amounts of materials sold for the various applications would be a reliable index to the suitability of the agents for the purpose. If this were so, it would be necessary only to ascertain which materials were most widely used for the different types of operations, and then to interpret the requirements of the operation in terms of the properties of the most popular agent. Several factors, however, tend to obscure the issue:

---

[1] Price, Donald. Ann. N. Y. Acad. Sci. 46(6): 407-424. 1946.

1. The intensity of the sales promotion of the various materials.
2. The physical form, ease of solubility, etc., of the agents.
3. Psychological aspects, such as color, odor, foaming power, etc.
4. Inertia. That is, in some instances, the same agent continues to be used because its action has continued to be satisfactory, although newly developed materials might be more suitable.
5. Availability. This condition is frequently present, but particularly so in time of war.

It is fairly obvious that the existence of complete literature describing the properties and potential uses of commercial surface active agents, coupled with well-planned advertising and good coverage by an alert sales organization, will tend to increase the acceptance of a particular material in comparison with comparable or even superior products which are less intensively offered.

It should be further realized that, after recommendations are made by the technical staff of an industrial user of surface active agents, matters are frequently in the hands of the purchasing agent and the production department. Thus, factors like the cost per pound, the type of container, the physical form, and the *apparent* efficiency, are of considerable importance. Some of the factors influencing the decisions of the purchasing agent or superintendent are primarily psychological. A very excellent material, at a low price, will seldom be purchased, if there is a tendency towards lack of homogeneity, as in the case of viscous fluids which stratify There is a pronounced preference for solids, either in powder or flake form, as opposed to liquids or pastes There seems to be a reluctance to buy a material containing water (which may be necessary in a final neutralization operation, for example) although drying to form solid materials may require the addition of a considerable amount of inert salts at a finite cost per pound, as well as the expense of evaporating water.

The foam problem, as well, is important. In some operations, it is necessary that there be no foam to cause overflow or floating of the material being processed In addition, the presence of large amounts of foam sometimes prevents intimate contact of the solution with the material being treated On the other hand, foam is sometimes of definite value. Even if the presence or absence of foam is not important as far as the efficiency of the operation is concerned, the operator (and his superior) generally feel that something is happening if they can see a considerable volume of foam on the solution. This is par-

ticularly true in the case of those agents which are being used in operations in which soap was previously employed.

It may be realized from the above that the relation of properties to suitability may not be thoroughly determined on the basis of the relative quantities of various agents used in a given operation. Some other method is therefore necessary for the determination of such efficiency.

If the use of a surface active agent is indicated, the available materials may be examined by pure trial and error; or the apparent factors involved may be studied and various possible materials checked on the basis of their physical properties. Either of these procedures presupposes the existence of a laboratory procedure which will serve to indicate the relative efficiency of the agents tested in respect to their actual use in practice. It is generally impractical to perform trial and error experiments in full commercial operation, and even if the physical factors contributing to efficiency are studied, it is almost always necessary to submit the agents which appear to meet the requirements to some sort of test for suitability. Except in the most simple cases, the laboratory testing of materials by methods corresponding to the industrial operation presents considerable difficulty. This is largely due to a tendency towards over-simplification of the testing procedure.

THE EVALUATION OF TEXTILE WETTING AGENTS

A good example of the tendency towards over-simplification is found in the evaluation of textile wetting agents. The wetting of textiles has been one of the most common uses of surface active agents and the methods of examination are so simple as to indicate that the testing is quite easy. Such, however, is not the case. Many methods, depending upon yarn shrinkage or upon absorption measurements have been suggested, but the two methods most widely used depend upon measuring the sinking time of textiles immersed in wetting agent solutions.

The canvas swatch method depends upon the sinking time of a small swatch of canvas placed upon the surface of the solution. A variation of the procedure[2] employs disks of canvas which are held barely submerged by means of a tapered glass tube or by other suitable devices. This leads to somewhat greater reproducibility than when the swatch is merely placed upon the top of the solution. When the swatch is placed upon the top of the solution, however, somewhat more information as to the nature of the wetting agent under test may be obtained.

[2] Seyferth, H., & G. M. Morgan. American Dyestuff Reporter 27: 525. 1938.

If varying concentrations of either sodium heptadecyl sulfate or the sodium salt of dioctyl sulfosuccinate are tested by placing canvas swatches upon the surface of the solution, it will be found that the wetting behavior of dilute solutions is somewhat different from that of dilute solutions of some other types of wetting agents. While relatively concentrated solutions produce rapid and thorough wetting and cause the swatches to sink without pause, the dilute solutions appear to wet the surface of the swatch some time before sinking is produced. In extreme cases, the top of the swatch will be completely wetted within five seconds and the swatch will then rest below the air-liquid interface for another thirty seconds before it sinks. It would appear that this behavior may be explained on the basis of the difference in rate of spreading on the surface of the yarn, as distinguished from penetration into the fine structure of the yarn.

Since the end-point in determination of wetting efficiency by the swatch method is the point at which enough water has been absorbed to cause the density of the cloth to be greater than that of the solution, it is apparent that a fairly large amount of penetration is necessary in comparison to the amount of penetration at the end-point of the Draves Test.

In the Draves Test,[3] a standard skein of cotton yarn is held submerged by the weight of a hook which in turn is attached, by means of a string, to a relatively heavy weight resting upon the bottom of the cylinder in which the experiment is being performed. Part of the buoyancy of the yarn is therefore overcome by the weight of the hook, and less water need be absorbed in order to produce sinking. It is obvious that, by changing the weight of the hook, the amount of absorption producing sinking may be varied at will. If the hook is almost heavy enough to cause the skein to sink in pure water, very low concentrations of wetting agents may promote displacement of air from the surface of the yarn and allow the skein to sink. If, on the other hand, the weight of the hook is so small that it barely holds the skein submerged, a very considerable amount of penetration must occur before sinking takes place, and the results become more comparable to the test in which the canvas swatch is held beneath the surface.

Experiments on a considerable number of commercial wetting agents indicate that the relative efficiency of the agents is quite dependent upon the weight of the hook used in the test. Frequently, the very fast wetting agents for sinking of skeins with heavy hooks are actually

[3] Assoc. Textile Chemists and Colorists Yearbook 21: 199. 1944.

slow wetting agents when light hooks are used. As might be expected, the agents which produce rapid wetting followed by slower sinking in the swatch test produce very fast sinking in the Draves Test, with heavy hooks, and proportionately much slower sinking when light hooks are used. It also appears that, in homologous series such as the straight chain primary sodium alkyl sulfates, the effect of increasing hook weight is less in the case of the lower molecular weight members of the series. The effect is best checked by adjusting the concentrations of the solutions of compounds under test until the sinking time is nearly identical with a hook of given weight, and then testing with a hook of a different weight.

By a consideration of the nature of the operation for which a wetting agent is needed, it may be decided whether a light or heavy hook is to be used in the evaluation by the Draves Test The next step is a decision as to the time in which this degree of wetting should take place Then, if the wetting time at various concentrations is plotted on logarithmic paper, the concentration corresponding to that wetting time may be obtained by interpolation from the resulting curve, which approaches a straight line Obviously, the relative economic efficiency of different agents is determined directly by considering the cost of solutions necessary to produce the desired degree of wetting in that time. There are still other factors which must be investigated before the final selection is made.

Textile operations are either batch or continuous in nature. In continuous operations, considerable quantities of fabric are passed through the solution. This may result, in some cases, in the adsorption of some of the constituents of the batch, as described by Shedlovsky.[4] On the other hand, foreign matter in the goods may accumulate in the solution and exert a considerable effect on the wetting efficiency. In addition, of course, it is sometimes necessary for the wetting agent to possess efficiency as a dispersing agent in order to prevent separation of insoluble foreign matter which has been removed from the goods Experiments should, therefore, be made, in which the solution is used repeatedly to wet fabric of the type for which it will eventually be used.

Much of the investigation of the effect of time upon properties of surface active agents has been made with relatively pure materials With some commercial agents, on the other hand, which may be either deliberate or unavoidable blends of several molecular species, the change of properties with time in dilute solution is very much more

[4] Shedlovsky, Leo. Ann. N. Y. Acad. Sci. 46(6) 443. 1946.

marked. It appears that this may be due to the change in state of solubility. When the concentrated material is dissolved in water, it is reasonable to assume that mixed micelles containing the several molecular species are formed. Upon standing in very dilute solution, however, equilibrium appears to be reached in which those materials having least tendency towards micelle formation are displaced by molecules of substance with a greater tendency towards micelle formation This may be checked by noting the difference in the properties of solutions of mixtures formed by mixing dilute solutions as compared with solutions obtained by dissolving concentrated mixtures of the substances.

In batch operations, the adsorption of constituents of the bath or the contamination of the bath takes place to a lesser extent. In many such operations, processing is started at low temperature and continued to high temperature. It is obvious that some thought must thus be given to the temperature range over which the agent is effective, unless a sufficiently high concentration is employed to result in wetting before the temperature is raised. In any event, commercial operation is frequently dependent upon the care exercised by a workman and temperatures of application are among the most common accidental variables The agent selected should, therefore, exhibit relatively uniform behavior over the temperature range in which it is to be employed The effect of temperature is too seldom taken into account in the testing of wetting agents

THE EFFECT OF TEMPERATURE ON THE WETTING PROPERTIES OF SURFACE ACTIVE AGENTS

In an attempt to oversimplify testing, tests for wetting efficiency by the Draves Test are frequently conducted at room temperature and the assumption is made that the relative effectiveness will be the same at some higher temperature of use. This conclusion is totally unjustified The shape of the curves obtained when sinking time is plotted against temperature depends entirely upon the nature of the material under test. In the case of an homologous series of straight chain primary sodium alkyl sulfates, for example, it will be found that, beginning at a low temperature, the sinking time will gradually decrease as the temperature is increased until a minimum sinking time is reached. In the case of the lower molecular weight members of the series, the sinking time will then definitely increase as the temperature is raised, and in the case of sodium lauryl sulfate, for example, the wetting effective-

ness at high temperatures is quite poor. In such a homologous series, the temperature of maximum effectiveness increases as the molecular weight of the compound increases. This effect is even more pronounced in the case of the commercial sodium alkyl sulfates, which contain some free fatty alcohol. In this latter case, the temperature of maximum wetting effectiveness coincides very closely with the melting point of the alcohol. Sodium diisopropyl naphthalene beta sulfonate, on the other hand, is very effective at low temperatures, but becomes very rapidly less effective at high temperatures. Thus, it is essential that wetting tests be conducted at the same temperature at which the compound would be used industrially.

OTHER FACTORS INVOLVING THE SELECTION OF WETTING AGENTS

Thus far, the following factors have been considered:

1. The extent of wetting required.
2. Time in which wetting must occur.
3. Effect of repeated use of solution.
4. Effect of variation of temperature.
5. Effect of age of solution.

The manufacturer of commercial wetting agents must investigate other factors, as well, in order to have sufficient information on behavior in various operations. Some of these are as follows:

6. The effect of concentration on wetting time.
7. The effect of pH on wetting time.
8. The effect of dissolved electrolytes on wetting time.
9. The effect of hardness of water on wetting time.

Thus, if the results of test by the use of four skeins are to be considered sufficient for each point on a curve, well over a thousand skeins are required to investigate the wetting behavior of a compound considered to be of interest as a wetting agent, since the above variables may not be considered as independent.

THE PRECISION OF THE DRAVES TEST

There has been much criticism of the Draves Test on the basis of its lack of precision. Probably four factors are primarily responsible for the apparent lack of precision:

1. The lack of uniformity of the cotton yarn used in the test.
2. The lack of uniformity of method of preparation of solution.
3. Measurements made at temperatures near that at which wetting time for the material changes rapidly.
4. Measurements made at concentrations near that at which the wetting time changes rapidly.

It must be remembered that a single shipment of standard skeins represents the output of a section of the plant in which they are produced, and it does not follow that they are from the same shipment of raw cotton. Thus, the behavior of the individual skeins of a carton may differ widely. Extreme instances of this have been noted. Thus, over the entire temperature range below the boiling point of water, the usual standard skein is not wetted by distilled water in twenty-four hours. Skeins have been tested, however, which, although apparently normal at low temperature, are wetted out within ten seconds at 80° C. by plain water. In fact, wetting agent solutions at 80° C. wetted out this particular lot of skeins more slowly than distilled water.

The effect of method of solution or thermal history is greater than sometimes supposed. In this laboratory, various commercial wetting agents have been dissolved by different procedures, and the wetting behavior noted. The methods of solution included the formation of 1% solutions by gentle agitation at room temperature, as well as the dilution of concentrated solutions formed by dissolving at high temperature. The behavior of the eventual solutions of identical concentration were quite different. This was particularly true in the case of commercial agents known to consist of mixtures of molecular species. As a result, a standard method of solution has been devised. A 10% solution is first formed at 70° C., which is then diluted with sufficient water at room temperature to give a 1% solution. Further dilutions are made from this 1% solution, and each dilution allowed to stand for fifteen minutes before testing. The reproducibility of the test is considerably improved in this manner. The same procedure is employed when other properties are to be measured.

THE EVALUATION OF EFFECTIVENESS IN DETERGENT OPERATIONS

We have seen that the difficulty of evaluation of wetting agents lies primarily in the interpretation of the results which are obtained by relatively simple procedures. The evaluation of surface active agents

for detergency in operations involving textiles, on the other hand, has been largely dependent upon setting up a proper test procedure. Here, too, the attempts to oversimplify have retarded development in the field.

Numerous simplified tests for the laboratory evaluation of detergent solutions have been suggested. The type of procedure most generally encountered involves the washing of a test swatch of "standard soiled" fabric. In typical instances, the test fabric is immersed in a dispersion of carbon black in organic solvent containing both vegetable and mineral oil. It is assumed that the carbon black is bound to the fabric by the oil, and that the ratio of oil to carbon black on the fabric, after washing, will always be almost the same as on the unwashed fabric. If this is true, the extent of removal of carbon and oil, obviously, could be followed by measuring the amount of either present. The common test procedure consists in measuring the reflectance of the fabric before and after washing, and then calculating the per cent increase in whiteness, which is termed the detergency of the solution.

There are four principal difficulties in this method, as generally performed:

1. The artificial soil probably simulates the type of soil encountered infrequently. That is, there are many types of soils and stains, and results obtained by the use of a single type of standard soil should be interpreted with extreme care.
2. Detergent systems have been examined which tend to remove carbon in much higher proportion than oil. Thus, there is a relatively large increase in whiteness, but very little removal of oil. It is entirely possible that the reverse type of specific removal can occur.
3. The conditions for the preparation of soiled fabric have been insufficiently studied. Thus, in a continuous strip of fabric, portions have been found which definitely differ from the remainder of the fabric in ease of removal of soil. Woodhead, Vitale, and Frantz[5] have taken this into account and proposed a statistical method for arriving at numerical expressions for the extent of removal. This procedure averages out the variations rather than preventing them.
4. The fabric is almost invariably too heavily soiled to give significant results.

---

[5] Woodhead, J. A., F. T. Vitale & A. J. Frantz. Oil and Soap 21(11,333). 1944.

The fourth difficulty deserves careful consideration. Unless a fairly heavy deposit of carbon is present on the fabric, considerable difficulty is encountered in observing changes in effectiveness. If, for example, the reflectance of the fabric is reduced five per cent by the application of a carbon suspension, it is difficult to obtain photometric readings distinguishing between the efficiency of two different concentrations of the same agent. If the reflectance is reduced by an amount corresponding to seventy per cent of the original reflectance, significant changes in reflectance may be expected to result from solutions of different concentrations, or at different pH values, or under any other variations in conditions. It is indeed possible, if relatively large numbers of observations are made on heavily soiled fabrics, to draw smooth curves showing the effect of variables upon the percentage increase in reflectance. In fact, an investigator may be encouraged, because the very good detergents do not remove all of the soil and the very poor detergents remove almost none of the soil, so that a fair degree of differentiation may be obtained. It is significant, however, that the good detergents still leave the fabric with a readily measurable amount of soil. In fact, the fabric, after washing under supposedly good conditions, may still be considerably dirtier than is the ordinary garment when sent to the laundry. It is entirely possible that the forces responsible for holding the last small amount of soil to the fabric may be somewhat different from those holding the more easily removed superficial soil. The measurement of detergency by the use of such heavily soiled fabrics, therefore, is not necessarily measuring the type of removal which is necessary in practice.

In laundry practice, the whiteness retention is an important factor In the case of poor whiteness retention, white fabrics, which have become lightly soiled and are repeatedly washed, gradually acquire a grayish shade which is much more difficult to remove than the ordinary superficial soil. Thus, a white shirt, for example, which is washed by a process giving poor whiteness retention, will eventually become almost unwearable.

There are very pronounced differences in the whiteness retention obtained by the use of different synthetic detergents. This will certainly prove a factor in the attempts to market such materials as soap substitutes for household use.

Although no single laboratory test for fabric detergency has proved adequate, steps are being made in the interpretation of results obtained by such methods. It has been found, in investigations of synthetic detergents for use by the armed forces under special field conditions,

that the only reliable test is actual operation in full scale equipment on the type of articles which are going to be washed in the field. Tests on standard soiled fabrics appear of use only as rather rough screening tests to limit the number of agents to be tested in field trials.

The existing tests involving standard soiled fabrics have been used to determine the effect of various alkaline builders on detergency[6-12] as well as other factors in the removal of soil. Thus far, however, no significant correlation has been shown between the removal of soil and the classical physical properties of the solutions.

THE RELATION OF PHYSICAL PROPERTIES OF SOLUTION TO DETERGENCY

Many hypotheses have been advanced to explain detergent action,[13] which is assumed to include wetting of foreign matter, dispersion of foreign matter, and deflocculation. These functions are, in turn, assumed to be dependent upon surface tension, interfacial tension, viscosity and even sudsing of the solution. Electrostatic properties of the systems involving the detergent solution and the surface to be cleaned have also been considered. Little progress has been made, however, in predicting detergent efficiency of a solution by the measurement of the physical properties of the solution. The difficulty, here, is probably not so much the lack of consideration of the factors involved as the lack of ability to determine the *relative* importance of the individual properties. It is natural to attempt to predict detergency by measuring these properties, for, generally, it may be said that solutions of organic detergents lower surface tension, lower interfacial tension, have some dispersing action, and produce foam. Yet there are many materials which exhibit some of these properties to a very marked degree, without proving to be efficient detergents.

A classical example is saponin, which exhibits the characteristic properties of surface active agents, while of almost no value in detergent operations. Many surface tension depressants, many foaming agents, many wetting agents, and many emulsifying agents are of almost no value as detergents. It is particularly significant that many materials which are of considerable value as emulsifying agents for oils

---

[6] Snell, F. D. Ind. Eng. Chem. 25: 1240. 1933.
[7] Cobbs, W. W., J. C. Harris & J. R. Eck. Oil and Soap 17: 4. 1940.
[8] Powney, J., & R. W. Wood. J. Text. Inst. 30: T157. 1939.
[9] Morgan, O. M. Can. Jour. Research 8: 429. 1933.
[10] Rhodes, F. H., & C. H. Bascom. Ind. Eng. Chem. 23: 778. 1931.
[11] Vaughn, T. H., & A. Vittone, Jr. Ind. Eng. Chem. 35: 1094. 1943.
[12] Carter, J. D. Ind. Eng. Chem. 23: 1389. 1931.
[13] Bell, P. H. J. Phys. Chem. 31: 801. 1927.

will not remove oil from textiles or other solid surfaces and, in turn, that some of the materials which are effective in the removing of oil are much less efficient in the production of oil emulsions.

It must be concluded that far more work must be devoted to the extent to which the various properties of detergent solutions are involved in detergent operations before prediction may be made on the basis of examination of the classical physical properties.

THE EVALUATION OF FOAMING AGENTS FOR INDUSTRIAL USE

As previously mentioned, the foaming properties of surface active agents are frequently of importance. In some instances, the presence of foam appears to have little significance other than the psychological effect upon the user. In many other operations, however, it is of extreme value. The influence of lather in washing hands or hair, for example, is easily demonstrated. Here, it is mechanically easier to perform the washing operation if a lather is present. It is, in fact, apparently the solution associated with the lather which is responsible for the resultant cleaning. Similarly, in the washing of rugs, it is desirable for a voluminous and durable lather to be produced, as decreased absorption of solution by the rug results when lather is used rather than an aqueous solution, with consequent ease of drying, and reduced damage to the rug. Also, the cleaning is accomplished by rotary brushes, and the cushioning effect of the lather prevents damage to the pile of the rug. In the washing of surfaces, such as milk storage tanks, the presence of a small amount of foam on the surface indicates the portion which has already been washed, and simplifies the washing procedure. In many textile dyeing operations, the continuous blanket of foam on the surface of the solution serves to immobilize small particles of foreign matter, which would otherwise be picked up by the fabric and result in specks or blotches, particularly when light shades are involved.

It is not surprising that the requirements for foaming agents in these and other operations may differ. In addition to the limitations of water conditions, dissolved material, pH, etc., which have been considered in the evaluation of wetting agents, other factors must be considered. Among other properties of interest are the ease of formation of foam, the resistance of the foam to outside influences such as strong agitation, the time stability of the foam, and the lowest concentration at which the foam is obtained. In addition, the amount of solution held by the foam is of particular interest in the washing applications

mentioned above, where cleaning is accomplished by solution carried by the foam.

Many procedures have been devised for measuring the frothing, foaming, or lathering properties of solutions of surface active agents. These methods fall into several classes:[14-21]

1. Direct aeration.
2. Light agitation.
3. Strong agitation.

Direct aeration methods, as applied, measure the film-forming properties of the solution and the time stability of the foam which is formed. If the gas-liquid interface resulting from the formation of a bubble beneath the surface of a solution of surface active agent is regarded in the customary light, it is apparent that the gas bubble is, in effect, surrounded by an oriented and concentrated film of molecules of the agent. If the tendency towards production of such an interface is very pronounced, and if there are lateral forces between the oriented molecules, every bubble formed may exist as a separate foam bubble after the surface of the solution is reached. If this is the case, all such agents will give identical volumes of foam and each foam bubble will be of the same size. If the efficiency of film formation is not perfect, some of the gas bubbles will break as they reach the surface of the solution, but may still be trapped by the mass of foam above the solution, thereby producing an increase in the volume of foam, but with variable bubble size. Foaming ability as measured by such methods is, therefore, not necessarily related to the amount of foam which would be formed from a solution of the same concentration in practice, unless the operation involves passage of a gas through the solution.

The production of foam by gentle agitation is a measure of the ease of formation of foam, as well as the time stability of the foam produced. No indication is obtained, however, of the durability of the foam under external mechanical force. Frequently, the amount of foam produced by gentle agitation is much greater than under vigorous agitation, due to the disruptive action of the strong agitation upon the foam bubbles after formation.

Ross and Miles[22] have considered this effect, and have devised a pro-

[14] Sier, A. Kolloid Zeit. 77 (1): 27. 1936: 78 (2): 156. 1937.
[15] Bikerman, J. J. Trans. Faraday Soc. 34: 634. 1938.
[16] Ostwald, W., & W. Mischke. Kolloid Zeit. 90 (1): 17. 1940.
[17] Foulk, C. W., & J. N. Miller. Ind. Eng. Chem. 23: 1283. 1931.
[18] Lederer, E. L. Seifensieder-Zeitung. 63: 331. 1936.
[19] Ostwald, W., & A. Sier. Kolloid Zeit. 76 (1): 33. 1936.
[20] Clark, G. L., & S. Ross. Ind. Eng. Chem. 32: 1594. 1940.
[21] Christensen, L. J. U. S. Pat. 1,[illegible]. July 5, 1932.
[22] Ross, J [illegible] Oil and Soap 18: 99. 1940.

cedure for the production of foam depending upon controlled turbulent pouring of solutions. The disruptive force of the liquid striking foam which has already been formed eliminates the more readily destroyed bubbles.

The minimum concentration for foaming may be measured by the gradual addition of a solution of surface active agent to water, with intermittent agitation. The smallest concentration required for production of a stable foam may be readily noted. If the agitation employed is similar to that encountered in practice, it is not difficult to differentiate between foaming agents on the basis of the amount required to produce an adequate volume of stable foam.

THE TENDENCY FOR FILM FORMATION

Foulk and Miller[23] have studied the tendency for film formation in gas bubbles beneath the surface of a liquid by observing the behavior of two bubbles allowed to come together after formation in the liquid. Trapeznikov,[24] by relating stability of single bubbles to results obtained by the use of the film balance on layers of water-insoluble polar compounds, has correlated film formation with effective molecular area of the oriented molecules and has found maximum stability of bubbles at an average molecular area of about 50 Å^2, i.e., when vaporous films are present. The same investigator has related bubble stability to surface viscosity, and has concluded that surface viscosity is the principal factor governing foam stability.

THE SIGNIFICANCE OF INDIVIDUAL PHYSICAL PROPERTIES IN INDUSTRIAL APPLICATION

It is well to look at the known significance of physical properties of surface active agents in respect to their industrial use. Unfortunately, there are very few clear-cut instances of direct correlation between the classical properties and the end-use suitability.

SURFACE TENSION

As surface active agents reduce the air-liquid surface tension, there is a pronounced tendency to regard the reduction of static surface tension as a primary property for study in connection with mechanism of use. The fallacy should be most apparent. In recognition of the effect of age upon the surface tension obtained, for example, by the du Nuoy

[23] Foulk, C. W., & J. N. Miller. Ind. Eng. Chem. 23: 1283. 1931.
[24] Trapeznikov, A. A. Acta. Physicochemica U. R. S. S. 13: 269. 1940.

instrument, ample time is generally allowed for equilibrium to be reached. This is in striking disregard of the fact that the systems in which such agents are employed are nearly always dynamic in nature, and the significance of static measurements conducted on well aged interfaces is hard to understand. The effect of temperature is also obscured, due to the difficulty of making measurements at elevated temperatures.

Not only does the measurement of surface tension by the du Nuoy instrument assume an undue significance for the static surface tension, but there is a strong tendency to relate efficiency to a low value of surface tension, although it has been clearly shown by Stamm[25] that, in the case of operations requiring the penetration of aqueous solutions into porous structures, the extent of penetration, as well as the rate of penetration may actually be less in the case of solutions of low surface tension. Thus, in the general equation for capillary rise: $h = \frac{2\gamma \cos \theta}{grd}$, it will be noted that, for a given value of θ, the contact angle, the lower the surface tension, the less will be the capillary rise. If the capillary wall is of such a nature that it is completely wetted by water, the addition of surface tension depressants will considerably reduce the extent of penetration. If the capillary walls are not wetted by water, the relationship of surface tension to the cosine of the contact angle is important, and increased capillary rise may be obtained only if the effect of the agent on the cosine of the contact angle is greater than its effect on the surface tension. The same relationship holds for the rate of capillary penetration, which may be represented by the general equation: rate $= \frac{\gamma \cos \theta \cdot r}{4\eta \cdot l}$.

It is apparent that, in solutions of low surface tension, the size of bubbles of gas formed at a given orifice will be less than in plain water. This fact has been advanced as an explanation for the effect of surface tension depressants in the reduction of pitting in electroplating, where the adhesion of gas bubbles to the metallic surface may be reduced.

INTERFACIAL TENSION

Emulsions are more readily formed if the interfacial tension between the continuous and disperse phase is low. Such ease of emulsification, however, does not guarantee emulsion stability, as the possibility for non-elastic collision of emulsion droplets still exists. Similarly, stable

[25] Stamm, A. J., & W. E. Petering. Ind. Eng. Chem. 32: 809. 1940.

emulsions may be formed in cases of higher interfacial tension, if sufficient mechanical work is done in reducing droplet size and the system is such that non-elastic collisions do not result. The relationships between choice of emulsifying agent and protective colloid in effecting emulsion stability are so complex as to discourage any but the most broad generalizations. Thus, it would appear that, whereas, according to classical theory in a two-component system, the ease of emulsification is inversely proportional to the interfacial tension, the multiple component systems encountered in industry resist positive prediction based upon interfacial tension.

Similarly, although repeated attempts have been made, interfacial tension has not been successfully related to detergency. We may strongly suspect that reduction of interfacial tension is of considerable importance, but the difficulty of independently altering a single variable in such a study prevents establishment of quantitative relationships.

THE SIGNIFICANCE OF CHEMICAL PROPERTIES IN INDUSTRIAL APPLICATIONS

In addition to the requirement for efficiency in particular ionic environments, as in hard water or in acidic solutions, other chemical properties of surface active agents are frequently the controlling factor in end-use suitability. The chemical stability, for example, should be considered before the agent is submitted to an extensive testing program.

It is well known that sodium alkyl sulfates are not resistant to prolonged boiling in strongly acidic solutions. Similarly, compounds which are esters of carboxylic acids hydrolyse readily in hot alkaline solutions. Such compounds include, for example, sodium dialkyl sulfosuccinate and the sulfonation products of monoglycerides, as well as some nonionic surface active agents obtained by the esterification of polyhydric alcohols or polyglycols with fatty acids. Unsaturated compounds frequently may reduce the efficiency of oxidizing agents or may be rendered ineffective when employed in conjunction with oxidizing agents. The same is true of some types of nitrogen compounds, particularly when used in solutions containing available chlorine. Incomplete or inaccurate information concerning the constitution of industrial surface active agents may lead to incorrect assumptions of stability in practice.

SOME ILLUSTRATIVE INDUSTRIAL APPLICATIONS

Some confusion exists in the descriptive terminology applied to surface active agents in practice. A material used as a detergent may be loosely classed as a "wetting agent;" penetrants may be spoken of as "synthetic detergents," etc. In many instances, several properties of the agent are responsible for the effect obtained. The following list presents a few of the more clear-cut applications. Wetting agents may be used for many purposes, where the effect is spectacular. Some of these are listed below:

The removal of dust from air by spraying wetting agent solutions.
Facilitation of fire-fighting, as in the penetration of cotton bales.
Prevention of pitting in electroplating.
Decreasing amount of water required in ceramic manufacture.
Insuring sharp meniscus in laboratory titrations.
Prevention of adhesion of air bubbles to photographic film during developing.
Improving the wettability of fabrics by drying after treatment with wetting agent solution. This backwetting is not dependent on the efficiency of the agent when used in the regular manner.
Improving the spreading efficiency of oils, lacquers, etc.
Admixture with insoluble solids to facilitate wetting by water.
Wetting of fabrics prior to further processing.
Wetting of particles of soluble materials, such as dyestuffs, to facilitate solution.
Removal of sand from lettuce, spinach, etc.

When functioning as emulsifying or dispersing agents, many types of emulsions may be prepared:

Oils which emulsify upon pouring into water. These include cutting and grinding oils, oil sprays, etc.
Emulsions of waxes in high concentration of aluminum ions.
Emulsions of resins which dry to clear films.
Emulsions with positively charged particles (As in the reversal of charge in latex emulsions).
Emulsions which may be exhausted, as in paper or textile processing.
Wet particle size classification.
Emulsion for polymerization of monomers under the influence of a great variety of catalysts.

For detergent operations, unique conditions may be fulfilled:

Detergents for use in sea water without added alkali.
Detergents for use in acid media.
Detergents for use on fabrics dyed with sensitive dyestuffs, or on surfaces painted with sensitive paints, etc.
Detergents for individuals with alkali-sensitive skin.

In addition to the above, there are other uses so well known as to make further mention unnecessary. It is a certainty that as more agents become commercially available, new uses will develop, and, finally, through a better understanding of the factors involved, an even greater efficiency will result.

CONCLUSIONS

Comprehensive studies of simple systems comprising solutions of pure surface active agents in water have been made by numerous investigators employing highly refined techniques. Much has been added to the knowledge of the physical interpretation of surface phenomena as a result of these studies. When investigations of this type have been extended to systems containing greater numbers of components, under the highly dynamic conditions encountered in most industrial applications, improved understanding of the mechanism of industrially important operations will be obtained. At the present, however, such understanding is qualitative rather than quantitative.

Annals of The New York Academy of Sciences
Volume XLVI, Art. 7. Pages 531–678
July 30, 1946

NON-PROJECTIVE PERSONALITY TESTS*

By

Harold A. Abramson, Keeve Brodman, Harold J. Harris, George G. Killinger, Bela Mittelmann, Zygmunt A. Piotrowski, David Rapaport, Roy Schafer, Martin Scheerer, David Wechsler, Arthur Weider, Harold G. Wolff, Edith Wladkowsky, and Joseph Zubin

CONTENTS

PART I: PERSONALITY INVENTORIES

PART II: THE CORNELL INDICES AND THE CORNELL WORD FORM

PART III: ABILITY PATTERNS AND PERSONALITY

PART IV: THEORY

* This series of papers is the result of a Conference on Non-Projective Personality Tests held by the Section of Psychology of The New York Academy of Sciences, March 30 and 31, 1945.
Publication made possible through a grant from the Conference Publications Revolving Fund.

NON-PROJECTIVE PERSONALITY TESTS

PART I

PERSONALITY INVENTORIES

THE EFFECT OF ALCOHOL ON THE PERSONALITY INVENTORY

By Harold A. Abramson

New York, N. Y.

It is a great pleasure to present to this Section the results of studies made with the Inventory of Hathaway and McKinley[1] during the past two years. In the pre-war days of leisurely, private practice of medicine, it was often possible to give as much time as necessary to the gentle exploration of the conflicts in our patients. However laudable and desirable an extended technic may be, a short method which leads to a useful psychosomatic history is of great value in peace-time, as well as in war-time, medicine. That such a short method may be found in the Minnesota Inventory is indicated by the data to be presented. Although this paper deals, primarily, with the results of experiments originally designed to study the effects of moderate doses of alcohol on the personality inventory, the clinical results have definite bearing on the validity of the test itself. The defects in the present tentative test procedure do not particularly invalidate our results. The way in which the test was used here was more or less self-controlling, because the same individuals were used repeatedly, with each individual serving as his own control.

SUBJECTS OF THE TEST

The experiments were conducted on a fixed army post. I was in the unusual situation of being the only medical officer in my Division, and of coming into constant contact with a group of officers who had advanced academic and engineering degrees. The wives of this army personnel, as well as the secretarial staff of the group, served as additional subjects. The average education of the group was, as a whole, unusually high. Understanding of the war effort was far above the average, because nearly all of the individuals served in some capacity connected with research and development. The group, as a whole, was keenly aware of the complexities of waging war and was deeply imbued with its responsibilities. Nearly all members of the group were moderate drinkers, or drank very little. No individual reported here is an excessive drinker. Indeed, it is of interest that, of the comparatively

large group accessible, with one exception, all of the individuals drank very moderately. Excellent *rapport* was established with nearly all of the subjects.

CONDUCTING THE TEST

Tests without alcohol were taken at odd times and places, as opportunity offered. Experiments with alcohol, however, were nearly always conducted in the same way. A standard dose of alcohol was administered in three cocktails, which were rapidly drunk between 4:30 and 5:00 p. m. in the afternoon, in a comfortable bar of the club house. Each cocktail contained 1.0 oz. of whiskey and 0.6 oz. of vermouth or sherry, so that a little less than 2 oz. of alcohol were rapidly administered. No food was permitted. The atmosphere was definitely social, and the environment, pleasing. In one or two instances, only two cocktails could be taken, because of the marked effect of the alcohol. In other instances, four cocktails were taken by those who were more used to drinking. In all cases, with certain deviations, the administration of the cocktails resulted in the level of alcoholic reaction sought: euphoria, talkativeness, slight unsteadiness and congeniality.

REACTIONS TO TAKING THE TEST

Refusal of the test was rare. Indeed, a study of the reasons presented by those refusing might be of interest. Most of those who took the test while sober were willing to take the alcohol test. Only a few of those who took the sober test refused to take the same test under alcohol, although perfectly willing to take a second test without alcohol Various pretexts were given for not taking the test under alcohol, such as "I don't like to drink," "I haven't the time now," "I want to wait a little while." In only one instance was the alcohol test refused because the person approached "never drank alcohol." It is of peculiar interest to me, and brings out a point emphasized in another paper,[2] that this person who "never drank alcohol" had the highest "lie score" of the male members of the entire group, including individuals not reported here.

REPETITION OF THE TEST

Hathaway and McKinley[1] discuss in detail the results of repeating the test in the same individual. FIGURES 1 to 4 and 35 are typical illustrations of what appears when the test is given (without alcohol) re-

peatedly to members of this group. The smooth curve in FIGURE 1 is the profile obtained in a subject recovering from combat fatigue. The results of this test were carefully explained to the subject, who had been a pre-medical student before the war, and who had become quite interested in the test itself. After some weeks, the subject decided that the profile first obtained did not represent his character traits and readministered the test to himself. In fact, he checked the records of the raw score and total scores of the first test administered by me. Observe the lowermost dashed curve in FIGURE 1. Whereas there is a change in the level of the Hy score, the general shape of the repeated curve remains essentially the same, with the F score increasing, in spite of a conscious attempt to lower the profile to a more acceptable level. Some months later, while the same subject was "angry, disgusted, demoralized, apprehensive and depressed," the test was taken again; the results being in the uppermost of the three profiles in FIGURE 1. The recurrence of the same general shape of the curve is noteworthy. In FIGURE 2, the smooth curve represents a sober test in a mature, intelligent, adult male. Some weeks later, this subject (voluntarily) asked for the cards again. He states that he had felt that he had only spent a short time taking the test, in the first instance, and that he would like to make a careful study of the cards to give a "true picture" of his profile. This subject retained the cards for several weeks and stated that, in the second trial, he "spent a good deal of time studying the cards before deciding on his answers." It was estimated that, instead of the usual 30–40 minutes taken for the first test, 5–10 times that period was utilized in taking the second test. The dots in FIGURE 2 depict the second test taken, three months after the first. Here not only the shape, but also the actual values agree in spectacular fashion. FIGURES 3 and 4 are illustrations of results of second tests, taken about one year after the first. The dots again represent the second test. Agreement both in actual values of the total scores and in the shapes of the curves is excellent. Compare FIGURE 4 with FIGURE 12. Note that the alcohol curve in FIGURE 12 more closely resembles the second sober run (dots in FIGURE 4). The same is also true for FIGURES 3 and 17, where the results of the repeated sober test agree more closely with the alcohol test itself. There was a period of some months between the alcohol and repeated sober tests in both cases. An insufficient number of cases of this type have been studied to warrant any conclusions that might be drawn from these two instances, in which the repetition of the second test more closely corresponded to the curve obtained after alcohol

THE EFFECT OF ALCOHOL

FIGURES 5 to 26 show the effects of alcohol on the personality profile. In all of these figures, the continuous lines are the results of the sober tests, whereas the dashed lines connecting the circles are profiles obtained after alcohol. Except for FIGURES 25 and 26, the tests are roughly arranged in the order of increasing divergence, in special categories between the sober and the alcohol curves. In FIGURES 5 to 8, the agreement between sober and alcohol profiles is very striking. There were two males and two females in this group. FIGURES 9 to 15 illustrate instances in which there were greater differences (10 or more) in the total score in 1 or 2 categories, but in all of these instances the general shape of the curve remains essentially the same. It is of some interest that, in FIGURE 9, the "F" score markedly dropped under alcohol, as may also be observed, to a less extent, in FIGURES 12, 15, 16, 20, and 22. In only two instances, FIGURES 17 and 23, did the "F" score rise after alcohol. Except for the curves illustrated in FIGURES 25 and 26, which I shall shortly discuss, there was no important change in the "cannot say" and "lie" score. FIGURES 16 to 24, as well as FIGURES 25 and 26, show a remarkable constancy in the shape of the curves following alcohol. Indeed, the agreement in shape also appears to be independent of the values of the total score. In an instance where the subject (FIGURE 21) had majored in psychology in college and apparently desired to have a score closer to the 50 line, an obviously conscious attempt on the part of the subject to be "normal" only resulted in a lowering of the curve as a whole, the shape of the curve remaining essentially the same.

In FIGURES 25 and 26, a special situation existed. These are inventories taken from the study of two sisters, 23 and 21 years of age, respectively. Note the similarity of the sober profiles to one another, as well as the agreement observed following alcohol with each of the sober runs. In particular, there is an unusual and dramatic elevation of the "lie" score. These are the only two instances in which the "lie" score was markedly elevated following alcohol. It is of interest to speculate if members of the same family, brought up together and of the same age, have a greater probability of giving the same personality inventories and of showing similar reactions to drugs by this type of test. It appears remote that chance alone was responsible for this elevation of the "lie" score. It seems likely that the elevation in both is connected with early training. There were twenty-two different subjects taking the alcohol tests. There was one chance in twenty-two that the first sister would have the elevation of the "lie" score. There was one

chance in twenty-one that the second sister would have the elevation of the "lie" score. The chance that the sisters should both show the elevated "lie" score is the product (21 × 22), or one part in 432. This indicates that it should be profitable to study families by the Minnesota Inventory. (I am indebted to Dr. Winsche for aid in the statistical treatment just presented.)

Differences of ten or more in the total scores occurred as follows: Hs:1; D:3; Hy:3; Pd:5; Mf·6; Pa:6; Pt:2; Sc:2; Ma:4. Because of the scoring method, no emphasis should be given to the quantitative aspects of this series, at this time. However, these questions may be asked: "What is the qualitative meaning of fluctuations after alcohol?" "Do these fluctuations provide leads to the resolution of unconscious conflicts?" More data, providing clinical correlations with the inventory, are needed.

NATURE OF RESPONSES AFTER ALCOHOL

The agreement observed between the profiles obtained after alcohol and the sober curves does not depend upon the identity of items selected for compilation of the raw score. Indeed, the number of consistent

TABLE 1

Figure	No. of Items Selected for the Compilation of the Raw Score		No of Consistent Selections	% of Column I Consistent Selections
	Total before Alcohol	Total after Alcohol		
5	130	137	94	72
6	75	79	36	48
7	49	38	25	51
8	46	35	21	46
9	112	93	60	54
10	71	88	52	73
11	65	65	26	40
12	107	108	58	54
13	144	139	103	72
14	91	132	70	77
16	128	96	62	49
17	99	75	50	51
18	151	116	84	55
19	127	113	83	65
20	164	219	130	79
21	176	152	124	71
22	87	102	55	63
23	126	180	114	90
24	62	104	44	71
25	96	103	40	42
26	104	114	62	60

selections of items for compilation of the raw score (TABLE 1) apparently bears no simple relationship to the agreement observed. Thus, in FIGURE 5, the number of items comprising the raw score before alcohol was 130 and, after alcohol, 137, with 94 items or 72% consistently selected after alcohol. In FIGURE 6, the percentage of consistent selections following alcohol is only 51, whereas, in FIGURE 11, only 40% of the items were consistently selected. TABLE 1 provides excellent evidence that the agreement observed between the sober and alcohol tests is dependent upon an inherent validity of the test itself, and not upon choice of the same items before, and after, alcohol, for the compilation of the raw score. This has also been pointed out in connection with the repetition of the tests without alcohol (FIGURES 1 to 4). TABLE 2 is a small part of a typical score sheet, and illustrates how varied the individual items scored without, and with, alcohol may be, although the fundamental shape of the curve is not affected. Thus, TABLE 2 discloses that the subject was "unable to tell everyone about himself" before alcohol, but could, after alcohol. He "loved his father" without alcohol, but denied it after alcohol. Indeed, the same difficulties in decision arise in regard to belief in a deity.

I have not had the time to study the variation in the way the items were answered before and after alcohol in specific cases, although the compilation of these answers is complete and ready for further examination. These contradictions or signs of indecision may lead to a method of clinically obtaining rapid information related to unconscious conflicts. As mentioned hitherto, the length of time required to take suitable psychosomatic history often makes the history so brief as to be of limited usefulness. The evidence thus far presented indicates that this test may be of value in the solution of that problem. Indeed, many of the subjects found that factors, which they had never previously thought important, appeared in a new and personal light, with emphasis upon situations evidently related to somatic and emotional difficulties.

THE CONSISTENCY PROFILE

I shall ask you to observe TABLES 1 and 2, which showed that, following alcohol, there were three main groups of items:

Group I. Items consistently scored without and with alcohol.

Group II. Items omitted for scoring after alcohol.

Group III. Items added for scoring after alcohol.

Group I, the consistent responses, have been treated in the following

TABLE 2

GROUP I.—*Items scored without and with alcohol*

A-3 I have never felt better in my life than I do now. (R)

B-54 My mother or father often made me obey, even when I thought it was unreasonable. (L)

C-2 Some of my family have quick tempers. (L)

C-6 At times, I have very much wanted to leave home. (R)

C-12 I have been disappointed in love. (R)

C-20 I am apt to pass up something I want to do when others feel that it isn't worth doing. (R)

C-27 It makes me feel impatient to have people ask my advice or otherwise interrupt me when I am working on something important. (R)

C-28 I find it hard to set aside a task that I have undertaken, even for a short time. (L)

C-48 I like to flirt. (R)

D-1 I never attend a sexy show if I can avoid it. (L)

D-2 I like to talk about sex. (R)

D-3 A large number of people are guilty of bad sexual conduct. (L)

D-10 I go to church almost every week (L)

D-11 I pray several times every week (L)

GROUP II.—*Items omitted for scoring after alcohol*

A-5 I do not tire quickly. (L)

C-21 I hate to have to rush when working. (L)

C-22 I have several times had a change of heart about my life work. (R)

C-42 I like to read newspaper articles on crime. (R)

D-37 I enjoy gambling for small stakes. (R)

E-10 I would certainly enjoy beating a crook at his own game. (R)

E-42 I refuse to play some games because I am not good at them. (R)

F-48 I never worry about my looks. (R)

E-51 I do not like to see women smoke (L)

F-17 I feel unable to tell anyone all about myself (R)

F-27 I am apt to hide my feelings in some things to the point that people may hurt me without their knowing about it. (L)

F-29 I do not try to correct people who express an ignorant belief. (L)

F-44 I have had periods of days, weeks or months when I couldn't take care of things because I couldn't "get going." (R)

GROUP III.—*Items added for scoring after alcohol*

B-13 I feel hungry almost all the time. (R)

B-21 I have never had any black, tarry-looking bowel-movements. (L)

C-15 I loved my father. (L)

C-30 I prefer work which requires close attention to work which allows me to be careless. (L)

C-31 I have at times stood in the way of people who were trying to do something, not because it amounted to much, but because of the principle of the thing. (R)

C-32 I usually "lay my cards on the table" with people I am trying to correct or improve. (L)

D-4 When a man is with a woman he is usually thinking of things related to her sex. (R)

D-15 I believe there is a God. (L)

D-30 At school I was sometimes sent to the principal for cutting up. (R)

D-44 I am always disgusted with the law when a criminal is freed through the arguments of a smart lawyer. (L)

way. These consistent items were scored as if they represented a raw score for the construction of a separate profile for each subject. I have named these curves, "consistency profiles" or "consistency curves." The broken lines connecting the dots, in FIGURES 27 to 48, are consistency profiles. A change was made in the way the subjects were arranged. FIGURES 27 to 37 are the males in the group. FIGURES 38 to 47 are female subjects. In general, in the calculation of the total score from the consistently scored items, the same or lower values are obtained in all categories, except in the "Mf" score for the female. FIGURES 38 to 47, therefore, will, usually, have the Mf score higher. Note, in all of the curves, that the consistency profiles (dashed line) have, in general, the same shape as the sober curves given for comparison. In FIGURE 30, the alcohol profile is also given for comparison. In this case, as in a few others, the shape of the consistency profile is closer to the alcohol curve. FIGURE 35 merits special discussion, for it illustrates a remarkable consistency in the shape of the profiles taken, at different times, over a period of two years. The dashed line is the sober curve. The uppermost unbroken line is the profile obtained during a very depressed state. The second smooth curve from the bottom is the alcohol curve, and the consistency curve is plotted in the same way as the rest in this group.

The full significance of the consistency profile can only be determined after a greater number of cases have been studied, and after the consistent items selected have been sorted As mentioned earlier, consistent items selected may well depend upon the drug employed. The data which I have presented, thus far, indicate that the method, as a whole, may be extended without difficulty, and with some certainty of fruitfulness, not only in the field of psychology, but also of pharmacology and clinical medicine.

To summarize:

1. The Minnesota Inventory provides a suitable clinical technic for obtaining fairly constant personality profiles, over extended periods, in an objective fashion. As such, it is an excellent method for rapidly obtaining a psychosomatic history.

2. It appears from limited studies that repetition of the test (without alcohol) may reflect a change in attitude, more by the variation in the height than by the shape of the curve. For this reason, it is believed that, even when within normal limits, the shape of the curve

may indicate the trend of future difficulties under adverse emotional stress.

3. Following the consumption of sufficient alcohol to produce moderate intoxication, the basic attitudes of the individual remain essentially unchanged in the special group of moderate drinkers studied here.

4. Even though the personality profile remains essentially unchanged, the consistent responses after alcohol represent only a part of the items chosen for scoring Other items are added, or omitted, after alcohol; usually without a marked change in the shape of the curve and with only minor changes in the values of certain categories. This indicates that there is an inherent validity, both in the group of questions comprising the test, and in the method of its administration.

5. Further evidence of the presence of an inherent validity of the test lies in the analysis of the "*Consistency Profiles.*" The consistency profiles are profiles constructed from the items answered in the same way before, and after, alcohol. The shape of the consistency profiles is similar to the profile obtained while sober or after alcohol, although fluctuations in special categories occur.

6 It would appear that studies of the effects of alcohol ingestion may be readily extended to different levels of intoxication and to different groups, especially alcoholics, in an objective fashion. An analysis of the special items answered differently before, and after, alcohol is planned. Investigations of fluctuations of this type may lead to new and simplified technics for more rapid resolution of unconscious conflicts.

REFERENCES

1. **Hathaway, S. R., & J. C. McKinley**
   1943. The Minnesota Multiphasic Personality Inventory Minneapolis University Press.
2. **Abramson, H. A.**
   1945. The selection of specialized military personnel. Psych. Med. **7**: 178.

CAPTIONS FOR FIGURES

FIGURE 1

The three curves illustrate the constancy in the shape of the curve at different times and under different conditions. The subject was an officer recovering from combat fatigue. The unbroken curve is the result of the first test. The lowermost dashed line is the second profile obtained when the first curve was consciously rejected and a better curve was sought. The third uppermost dashed curve was obtained while he was "angry, disgusted, demoralized, apprehensive and depressed." Note that the validity (F) score went up for both of the dashed curves.

FIGURE 2

The smooth curve is the first test of a fairly well oriented person. The dots are the result of a second test taken some months later with more careful and lengthy scrutiny of the questions comprising the test.

FIGURE 3

The smooth curve was taken one year before the points were scored. The second curve is closer to the alcohol run. (See FIGURE 17.)

FIGURE 4

The points were obtained eight months after the smooth run and again are closer to the alcohol curve. (See FIGURE 12.)

FIGURES 5–8

These four curves illustrate the excellent agreement which may be obtained sober and after alcohol. The smooth curves were run without alcohol. The circles connected with dashed lines are with alcohol. In no instance is the difference between any two points in the categories as much as 10 Note that the center of gravity of the curves is near the fifty line.

FIGURES 9–24

In this group of figures, greater differences occur in certain of the categories. However, the alcohol curves approximate the sober curves. Note that, even though the center of gravity of certain of the curves is elevated in this series, as in FIGURES 13 and 15, the agreement of the shape of the curves obtained with and without alcohol is excellent.

FIGURES 25–26

Two of the 22 cases studied (two sisters) represent the only instances in which there was marked elevation of the "lie score" It can be shown that there is one chance in 432 that this is accidental. (For further discussion, see text.)

FIGURES 27–37

Consistency Profiles (Males). The effect of plotting the consistent answers only are given in the profiles by the dots connected with a dashed line. In FIGURE 30, the alcohol results (circles) agree more closely with the consistency profile. FIGURE 35 is of special interest. (For description, see text.)

FIGURES 38–47

Consistency Profiles (Females). Note that with female subjects the "Mf" score is apt to be higher than the sober value. (For further discussion of the consistency profile and its significance, see text.)

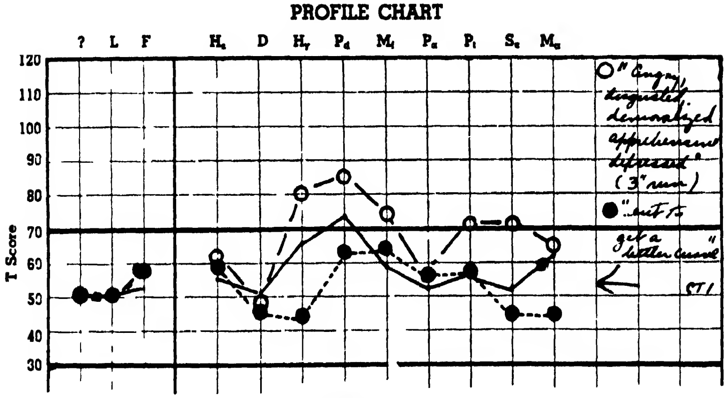


FIGURE 1

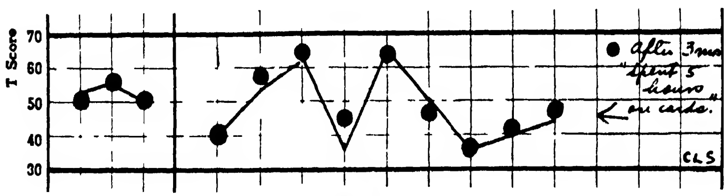


FIGURE 2

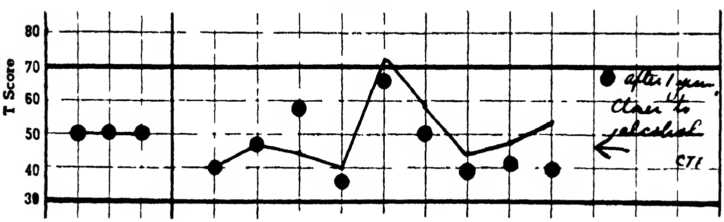


FIGURE 3

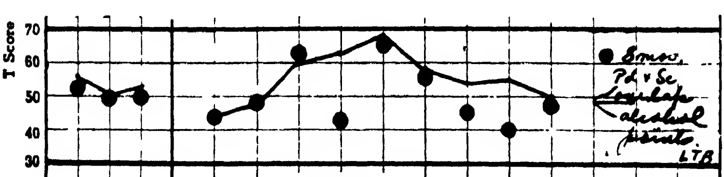


FIGURE 4

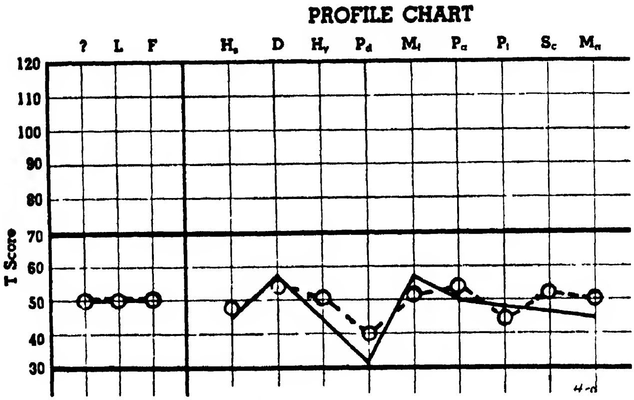


FIGURE 5

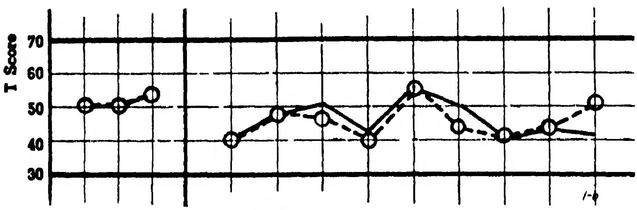


FIGURE 6

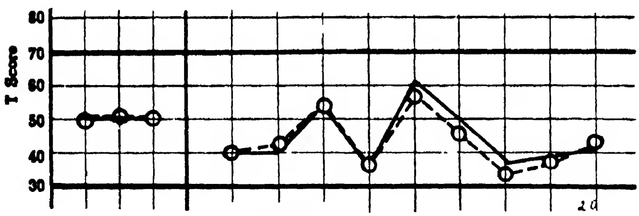


FIGURE 7

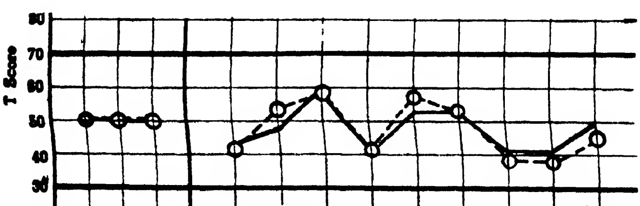
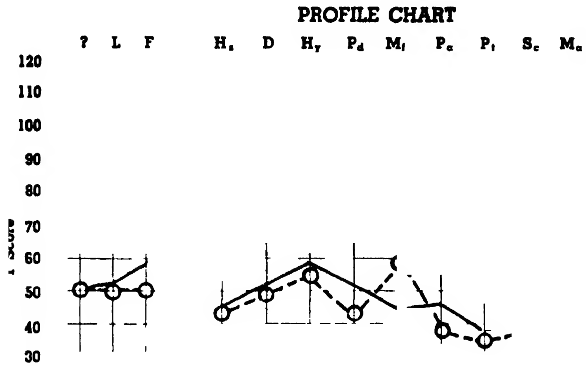


FIGURE 8



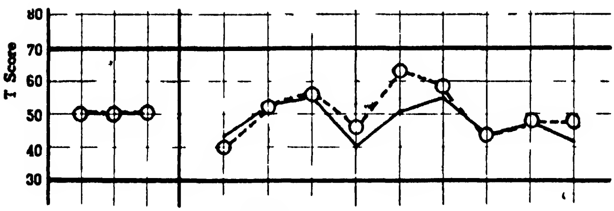
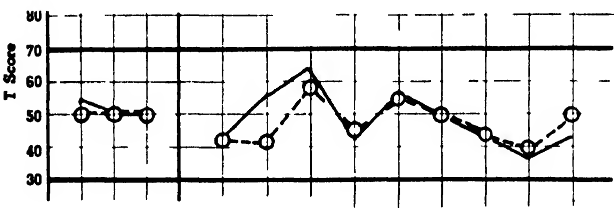


FIGURE 10



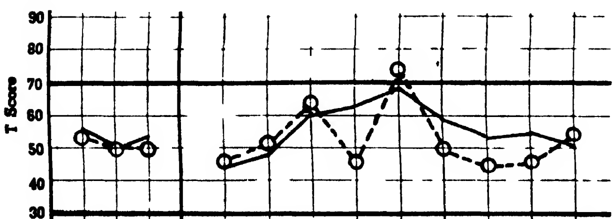


FIGURE 12

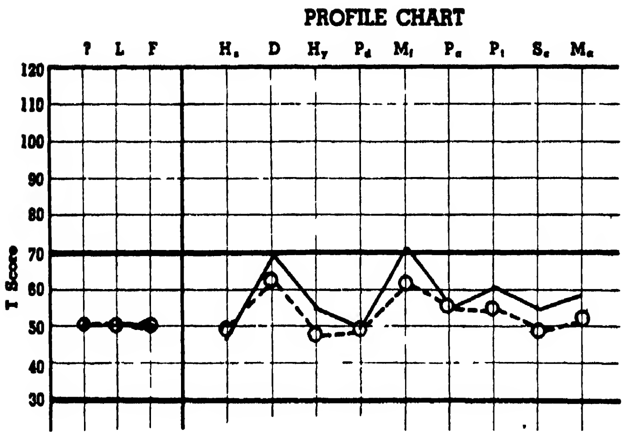


FIGURE 13

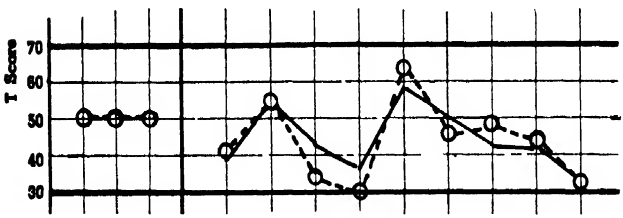


FIGURE 14

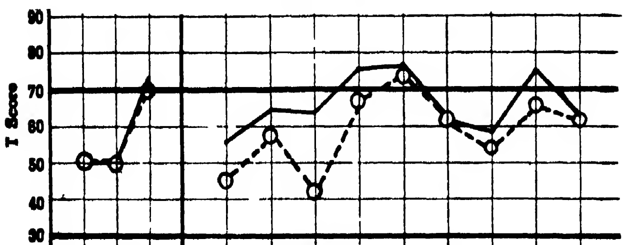


FIGURE 15

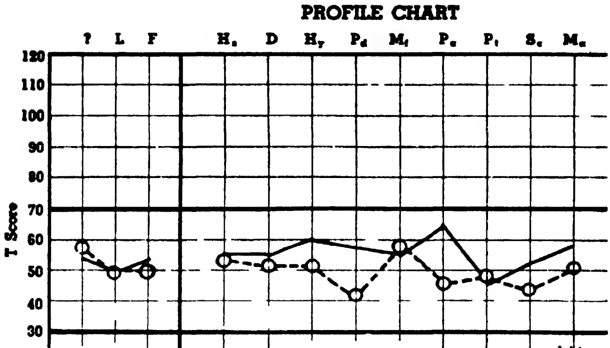


FIGURE 16

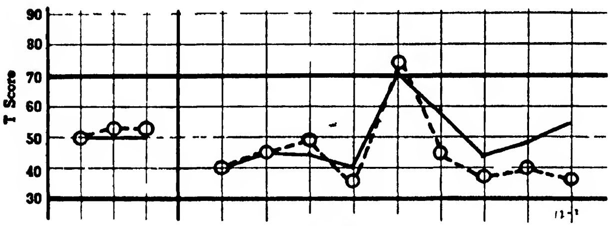


FIGURE 17

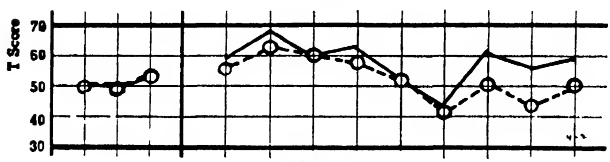


FIGURE 18

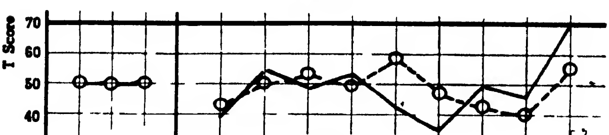


FIGURE 19

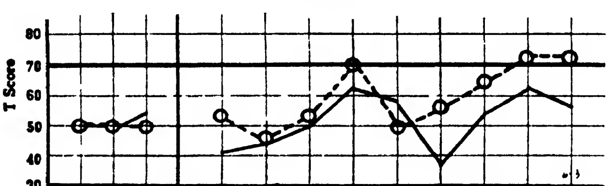


FIGURE 20

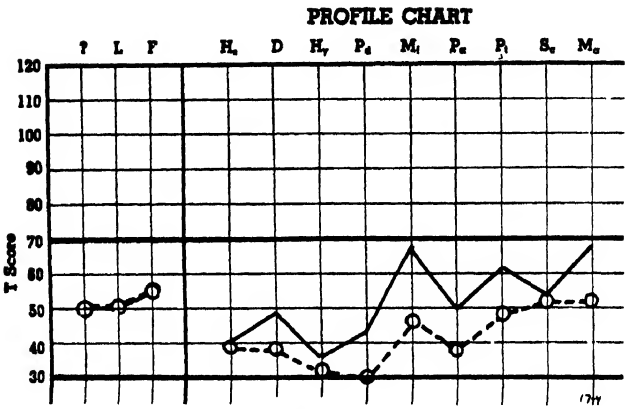


FIGURE 21

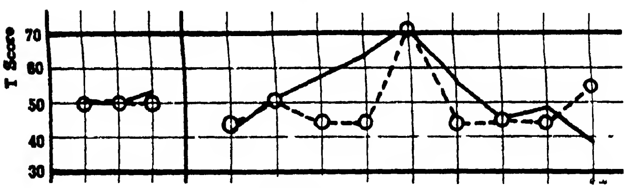


FIGURE 22

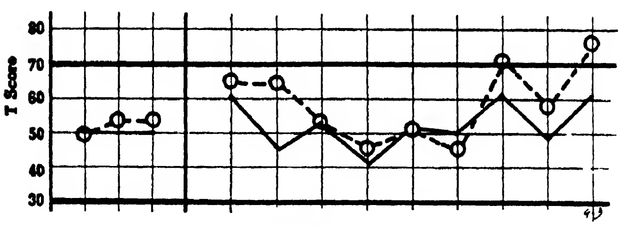


FIGURE 23

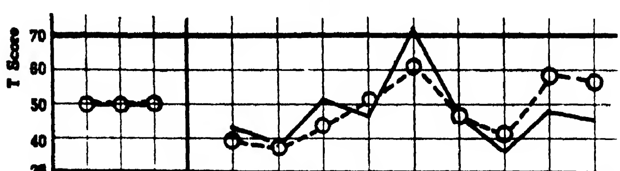


FIGURE 24

PROFILE CHART

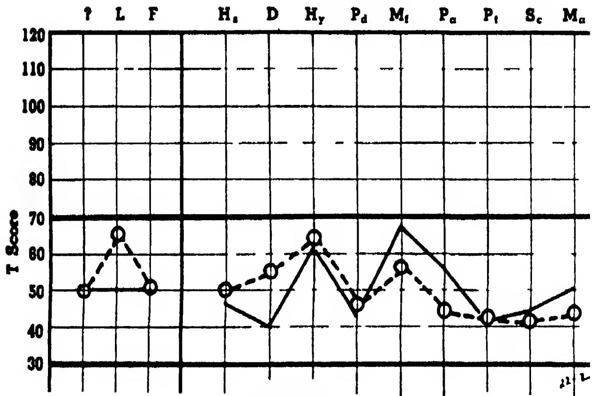


FIGURE 25

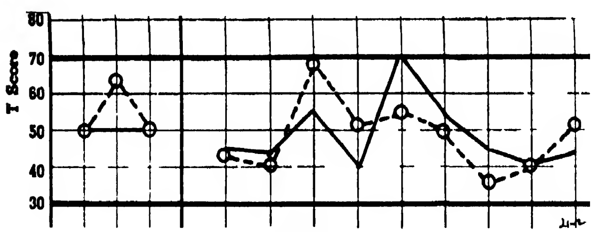


FIGURE 26

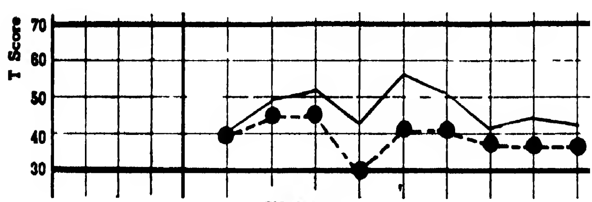


FIGURE 27

PROFILE CHART

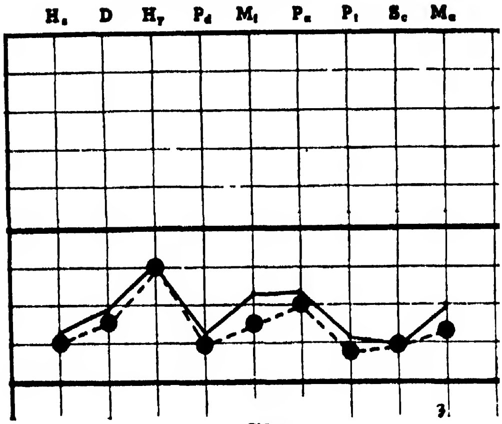


FIGURE 28

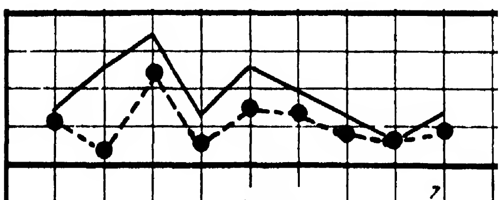


FIGURE 29

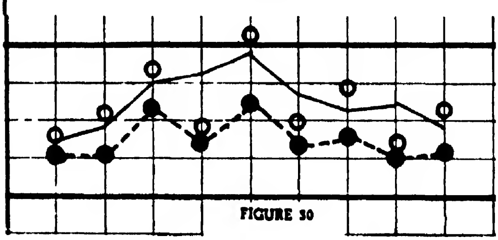


FIGURE 30

PROFILE CHART

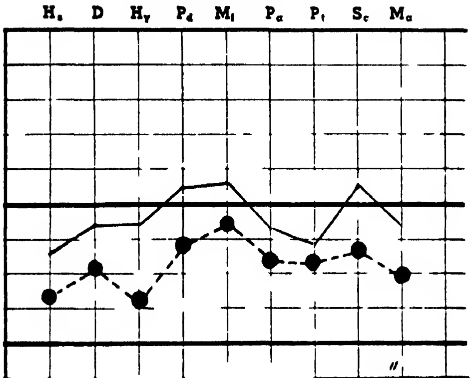


FIGURE 31

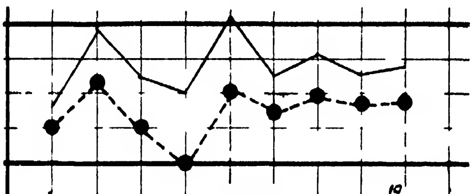


FIGURE 32

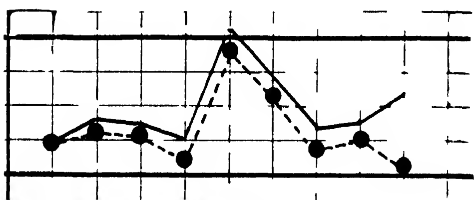


FIGURE 33

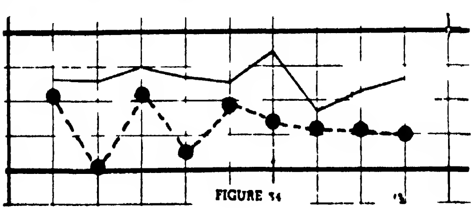


FIGURE 34

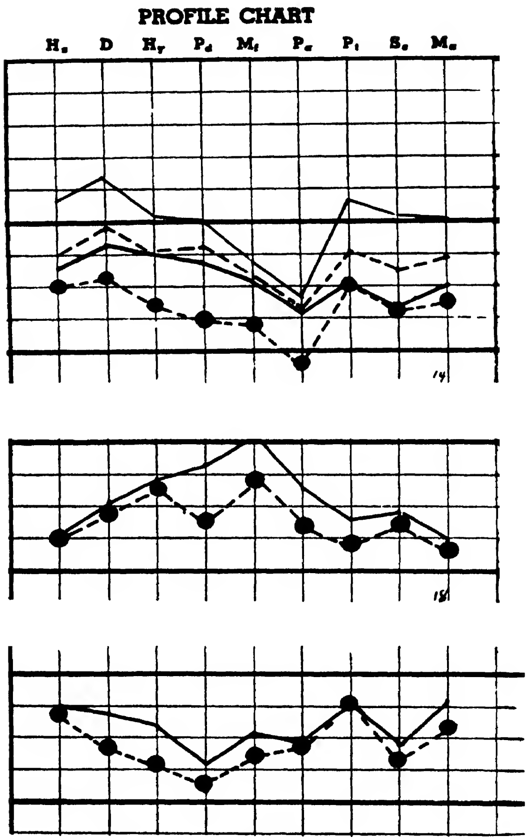
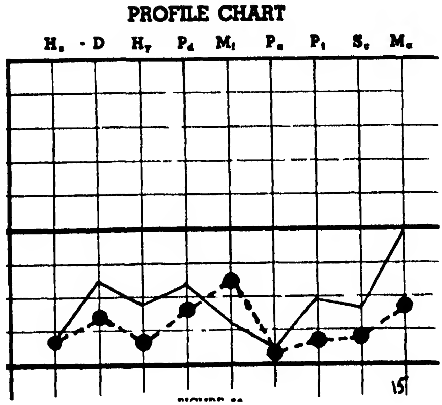


FIGURE 37



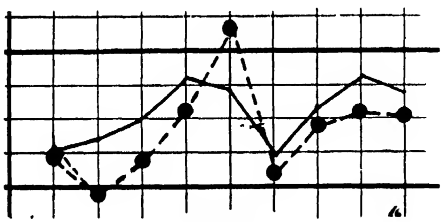
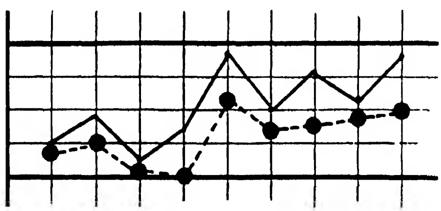


FIGURE 39



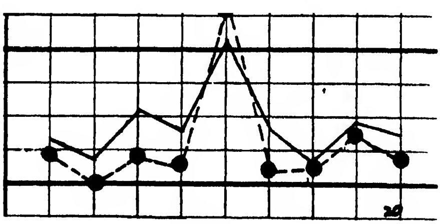


FIGURE 41

PROFILE CHART

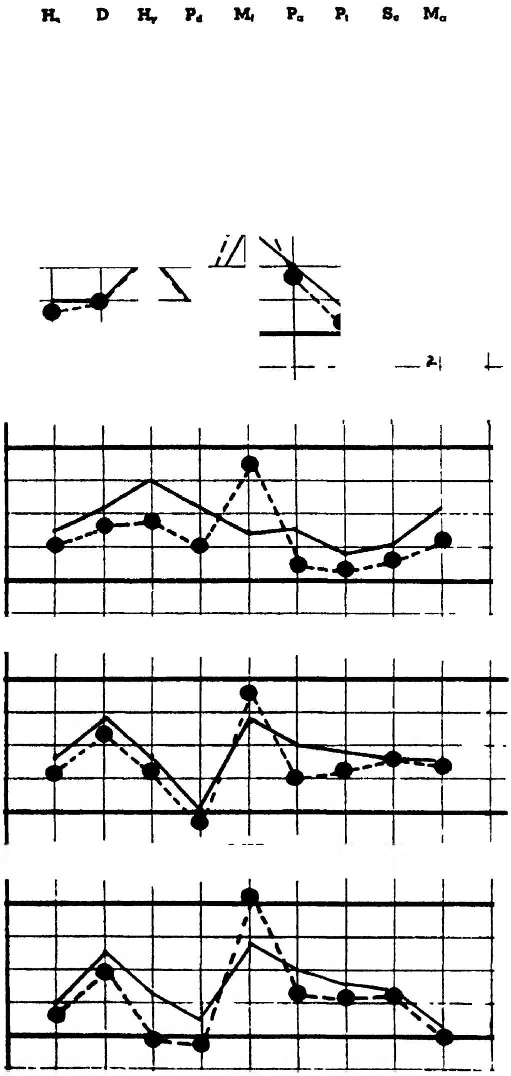


FIGURE 45

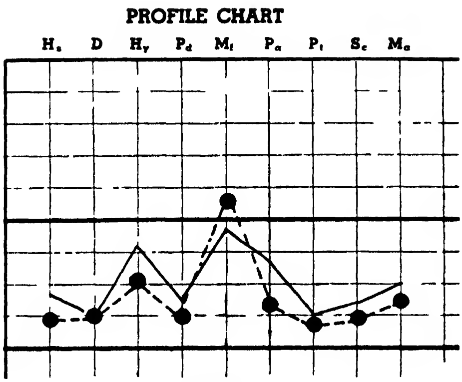


FIGURE 46

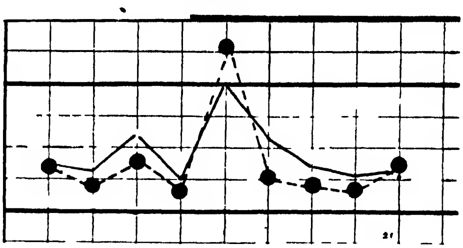


FIGURE 47

PSYCHOBIOLOGICAL SCREENING PROCEDURES IN THE WAR SHIPPING ADMINISTRATION*

By George G. Killinger† and Joseph Zubin‡

The War Shipping Administration was created, soon after Pearl Harbor, by Executive Order, as a temporary war-agency to control and operate American merchant ships. Approximately 90 per cent. of all military and essential cargo moved during the war in American vessels has been under the control of the War Shipping Administration, the remainder being controlled by the Army and Navy. In carrying out the responsibility entrusted to it, the War Shipping Administration had to cope with the problems of obtaining the necessary ships, manning them, planning cargoes, and signing on merchant seamen for voyages and signing them off at the completion of each trip. Of the many tasks which the War Shipping Administration had to face, the one that we are primarily concerned with was that of securing men to man the ships.

It should be realized that the American Merchant Marine is essentially a civilian organization, which, at the beginning of the war, consisted of some 55,000 experienced seamen in active service, plus a considerable number on the inactive list but, temporarily, out of the industry and in other employment, yet who, nevertheless, were potentially available to the industry. With the development of the war, the manpower needs of the industry have jumped to approximately 250,000 men needed for active service. The task of recruiting enough personnel to man an industry whose manpower requirements increased five-fold, presented selection problems of a psychological, medical, and vocational nature, which could not be met by the peace-time organization of the industry.

In order to cope with the problem of training new men and bringing back former seamen who had left the industry during peace-time, two separate divisions were created: (1) The Recruiting and Manning Organization, established to recruit experienced personnel; and (2) The Training Organization, to recruit and train inexperienced personnel.

Both of these organizations were faced with the necessity of obtaining psychological, psychiatric, and medical services for active seamen, as well as for the newcomers to this industry.

* Read before the Section of Psychology of The New York Academy of Sciences, March, 1945
† U. S. Public Health Service Reserve, Chief, Psychobiological Activities, WSA.
‡ U. S. Public Health Service Reserve, Chief Medical Statistician, Psychobiological Program.

In order to meet this need for a specialized war-time medical service, advantage was taken of the Federal statutes, dating back to 1798, which authorized the U. S. Public Health Service to provide medical care for American Merchant Seamen. Through all these years, the U. S. Public Health Service has continued to furnish medical care to Merchant Seamen through the operation of Marine Hospitals and Relief Stations. However, this service had neither been greatly concerned with a mandatory, pre-employment examination for each voyage, nor with the treatment of seamen suffering from war—and other types of neuroses. With the greatly expanded organization, existing facilities proved inadequate, in these particular respects at least. In order to meet the new needs, an "Office of the Medical Director" was created in the War Shipping Administration, to be filled by a commissioned officer of the Public Health Service. Later, the President of the United States recognized the significance of this work by promulgating an Executive Order which advanced the rank of this post to Assistant Surgeon General of the U. S. Public Health Service. Dr. Justin K. Fuller, who had served as Medical Director of the War Shipping Administration since the creation of that office in September 1942, continued in the new office. The medical staff of the War Shipping Administration is composed of four types of personnel: (1) Commissioned Officers of the U. S. Public Health Service, all of whom are detailed to the Navy Department and assigned to the War Shipping Administration; (2) Commissioned Medical Officers of the U. S. Navy; (3) technical personnel in commissioned and non-commissioned ratings in the Maritime Service; and (4) U. S. Civil Service Personnel.

The Psychobiological Service is one of the special professional services of the Medical Division of the War Shipping Administration, and was so named because it was desired to bracket under one heading that part of the medical program which dealt with clinical psychology and clinical psychiatry, in which the psychiatrist and psychologist work as a closely coordinated team.

This Service has the four-fold purpose of screening out the unfit, providing short-term therapeutic treatment for the mildly maladjusted who offer good prognosis, teaching normal psychology and methods of adjustment to the general body of trainees and seamen, and giving general psychological and psychiatric counselling to individual normal seamen, in order to equip them to meet the exigencies of sea duty during war time.

In addition to the psychological and psychiatric activities within the Psychobiological Service, the War Shipping Administration, through Sr. Surgeon (R) Daniel Blain, the Medical Director for the Recruiting and Manning Organization, who also serves as Medical Director for the United Seamen's Service, carries on an intensive psychiatric program in the selection and treatment of seamen who show evidence of occupational stress, such as convoy fatigue, war neurosis, and other pathological symptoms. The psychiatrists of this Service work very closely with the Psychobiological Service and serve as psychiatric consultants to the Medical Examination Program for seamen. This Organization has established, on the East, Gulf, and West Coasts, informal, non-institutional types of rural residence-units called Rest Centers, each under the direction of a psychiatrist. These Centers provide care for all seamen requiring a brief therapeutic regime of rest and recuperation. As Dr. Blain has so well pointed out, the aims of this brief therapeutic program are to build the seamen up, eliminate their symptoms, give them psychological security and prophylaxis against the hazards of subsequent voyages, and send them back to sea.

This paper is devoted to the description of only one part of the psychobiological program: that of screening. We shall deal, first, with the Training Organization. The Training Organization absorbed all the previously existing government training facilities for service in the Merchant Marine and expanded them. The present facilities consist of the U. S. Maritime Service and the U. S. Merchant Marine Cadet Corps. The Training Organization also extends supervision to the State Maritime Academies located in California, Maine, Massachusetts, New York, and Pennsylvania. The U. S. Maritime Service is the largest Unit of the present Training Organization, and it was with this Unit that the Psychobiological Activities were first developed and have been used most extensively. The Maritime Service operates: (1) Enrolling Offices, located in the major cities throughout the United States, whose purpose is to recruit applicants for apprentice training; (2) Training Stations for apprentice seamen, located at Sheepshead Bay, Brooklyn, New York; St. Petersburg, Florida; and Avalon, Catalina Island, California; (3) Training Stations for Officer Candidates who are experienced seamen, located at Fort Trumbull, New London, Conn., and at Alameda, California; (4) Upgrade Schools, in the larger seaports, for further training of cooks and bakers, able-bodied seamen, and officers; (5) Radio Schools, at Gallups Island, Boston Harbor; and Hoffman Island, New York Harbor; (6) Purser-Hospital Corps School;

(7) Turbo-Electric Schools and other schools for specialized training; (8) Training Ships, of which there are ten of major size, that are utilized in connection with the various training programs to give practical experience to the trainees. In addition, the Training Organization operates three basic schools for the training of cadet midshipmen: at Pass Christian, Mississippi; San Mateo, California; and King's Point, New York. It also operates the Merchant Marine Academy at King's Point, New York.

The screening program was inaugurated at the three Training Stations which devote the major part of their time to the preliminary training of apprentice seamen, and extensive screening procedures were established at each. It was soon found that preliminary screening of the obviously unfit was required at the source of enrollment (the Enrolling Office), in order to avoid needless expenditure in transporting the patently unfit from their homes to the Training Stations, only to be disenrolled and sent home immediately upon arrival. Such screening had already been instituted, at the Enrolling Offices, for physical defects, but the psychologically untrained personnel in charge of these enrolling offices could not readily detect the emotionally unfit. To meet this situation, two personality questionnaires were prepared; one for enrollee candidates under age 18, and another for candidates 18 years of age and older. Both of these questionnaires were based on questions selected from the more elaborate Personal Inventories used at the Training Stations, and from certain social history data, such as record of incarceration, 4-F classification by draft boards, and previous service in an armed force; all of which had been found significant in elimination of persons incapable of surviving Training Station routine and subsequent adjustment to sea-life.

Visits were made to each Enrolling Office in the major cities throughout the United States, and medical officers and pharmacist's mates were instructed in simple screening techniques. The screening at Enrolling Offices must necessarily be at a simple level, since the personnel consists, very often, of either pharmacist's mates or Medical Officers without much psychological or psychiatric training. Following the inauguration of these questionnaires at the Enrolling Offices, rejections at these offices perforce increased and rejections at the Training Stations dropped, with a net result that there was introduced into the training group a generally higher type of individual.

The Training Station screening procedure is as follows: The trainees who survive the screening at the Enrolling Offices, immediately upon

arrival at the Training Station and before receiving their physical examination, are assembled in large groups, and are asked to complete a Personal Inventory of the forced-choice type of some 80 items. This is supplemented by a short personal history sheet consisting of some 30 items built around family history, educational history, occupational history, and a self-evaluation by the trainee of his present state of health, together with a listing of any previous illnesses, injuries, or operations. The enrollee is also asked to state why he enrolled in the Maritime Service, in an attempt to gain insight into his motivation for entering the Service.

The trainees are introduced to these questionnaires by the statement that this is the beginning of their medical examination, and that, were sufficient personnel available, each trainee would be individually interviewed. Since this is impossible, they are asked to cooperate by filling in these forms as carefully as possible. In this way, *rapport* is established, and sufficient motivation aroused for careful completion of the forms.

As a result of some preliminary experiments at the Station, it has been found desirable to select for personal interview all trainees who make a score of 20 or more on the test; that is, those who give 20 or more abnormal answers out of the 80 items presented. In addition, several items have been selected from the Inventory which are themselves sufficiently indicative of probable maladjustment. These are items dealing with headaches, fainting spells, head injuries, dizziness, enuresis, arrests and imprisonments, a visit to a doctor or hospital for nervousness or convulsions. These are termed "stop" items, and a trainee who gives a deviant response to even one of these "stop" items is called in for personal interview, regardless of his total score. The personal background data are also scanned for possible implications of mental deviation. The following items in the personal history are considered sufficiently indicative to warrant a personal interview: discharge from military service (in the present war and prior to the termination of the war); arrests or imprisonments; history of mental illness, head injury, sleepwalking, and educational retardation.

The scores are computed immediately after the questionnaires are administered, and the results are used in selecting candidates for interview by the psychobiological team stationed on the main examination line. The men so selected are given a short interview by the psychologist or psychiatrist to determine whether they are grossly abnormal, mildly abnormal, or essentially normal, but had been accidentally caught

in the screening net, *i.e.*, the "false positives." These initial interviews usually last approximately 4 minutes. The grossly abnormal are placed on immediate recall by the Psychobiological Unit, which studies the case quite intensively, utilizing for this purpose the psychometric scores supplied by the Classification Division, neurological findings made by the psychiatrist of the Psychobiological Unit, physical findings supplied by the Medical Department, and other psychological tests of a clinical nature administered by the psychologists of the Psychobiological Unit. The mildly deviant individuals who show some mental deviation, but who do not give convincing evidence of being unable to succeed in training and sea duty, are asked to return for follow-up therapy within a week after beginning training. These men are followed for several weeks and, in some cases, eventually make an adjustment. In other cases, they are subsequently disenrolled. The "false positives," along with the normals, enter training immediately and receive only the routine group mental hygiene.

At the inception of the program, an attempt was made to find the most suitable Personality Inventory for screening purposes. After intensive research and experimental tryouts, the Personal Inventory, Format B, prepared by the National Defense Research Council, was selected because it had been validated on men in the military forces, and because of its forced-choice type of item which was believed to be more suitable than a more direct form of question for eliciting accurate information. After a preliminary trial of this Inventory on some 10,000 cases, it was found that the instrument was too unwieldy for our purposes. There were too many items that did not differentiate between successful and unsuccessful trainees. Some of the words used in the items were too difficult, and it was not especially suited for selection of men for the maritime industry. In order to circumvent these difficulties, an item analysis of the results on a sample of some 1400 men was made, and a thorough analysis by age groups as well as by cause of disenrollment indicated that some 100 items out of the 145 included in the Inventory proved to be diagnostic of failure at the Training Stations. It should be pointed out that some of the items which were found to be diagnostic had not been used diagnostically by the N. D. R. C. Inventory, but had been introduced merely as psychological ballast for the questionnaire. On the basis of the results of this item analysis, a new Inventory was drawn up which included 26 items taken directly from the N. D. R. C. Inventory, 21 modified items and, 33 new items, covering special Maritime situations, such as fear of water, high places, jump-

ing and climbing, and the general area of patriotism and motivation for joining the industry. It is well to recall, at this point, the essentially civilian character of the Maritime Service. Because the trainees are civilians who can come and go practically at will, the responses of these men to the questionnaire do not suffer to the same degree from the handicaps that might arise in the Military Service where a greater tendency may exist to put one's "best" or "worst" foot forward. Furthermore, because of their civilian freedom, it is important to select only men who are emotionally suited for the particular hazards that beset the seamen, especially in time of war, since there is no way of forcing these men to continue in the Service, if they find themselves unsuited for it. Whatever excellence the screening program achieves redounds, not only to the eventual welfare of the men themselves, but also to the eventual efficiency of the industry.

We may now turn to the description of the actual results of the screening procedure. As evidence of its efficiency, we summarize as follows: Out of every 100 men, 66 present evidence of being sufficiently emotionally stable not to require any further screening than is afforded by the written Inventory and Personal Data Sheet. The remaining 34 require, at least, a brief screening interview, because their Inventory and Personal Data Sheets indicate that they have either given more than 20 deviant responses, have given one or more deviant "stop" responses, or have given an indication in their personal history of some condition which requires further investigation. Of these 34 interviewees, only about 1 or 2 are found to possess personality defects making it mandatory that they be immediately excluded from training Some 6 or 7 are found to have temporary, or mild, emotional conditions, which, after a short stay at the Station under the guidance of the Psychobiological Unit, become sufficiently ameliorated to permit them to continue training. The remainder are individuals who have either misunderstood some of the questions, or have unintentionally given indications of conditions which, if they actually existed, would disqualify them. Thus, within one or two days after the arrival of the group, the vast majority of the misfits have been disenrolled, while those requiring therapy are given their follow-up appointment schedules. In the course of the first month of their stay, a few more individuals are found to have conditions or defects which require their elimination from the Station. Among these are included a small group of individuals who misrepresented themselves either intentionally or unintentionally in their responses to the written part of the screening procedure. These are the

enuretics, somnambulists, epileptics, and others who refrain from indicating their true condition for one reason or another. In addition, there may be several individuals who develop severe emotional disturbances and even psychotic episodes after their admission to the Station. These, of course, could not have been detected on admission, since the disqualifying condition was probably not present, or in remission, at the time.

Summarizing the efficiency of the screening procedure, it might be indicated that it successfully screens out fully 85 per cent. of those who eventually must be disenrolled. Some 12 per cent. who are eventually disenrolled escape the initial screening net through misrepresentation, while 2 or 3 per cent. develop personality disturbances, after their admission, of sufficient magnitude to require their disenrollment. It is hoped, eventually, to reduce the number of individuals who get by the screening procedure through misrepresentation, by utilizing various approaches to pattern analysis of the item responses.

It should be noted, at this point, that even those who are disenrolled are not summarily dismissed, but are given a careful exit interview which aims to prepare them, emotionally and vocationally, for finding a better outlet for their patriotic motives.

Screening does not stop at the Enrolling Office or the Training Station, but continues throughout the active career of the seaman. A special Wartime Order requires that each seaman receive a sign-on examination before each voyage to determine whether he is mentally and physically capable of making the trip. As part of this examination, a special psychobiological screening technique has been introduced. The screening questionnaire used consists of some 20 questions and deals chiefly with traumatic episodes such as torpedoings and air raids, as well as hospitalization overseas, treatment at Rest Centers, and special problems which may have developed for the seaman, since leaving the Training Stations and while employed in the Merchant Marine. Upon completion of these questionnaires, the seamen are interviewed very briefly by a psychologist on the Sign-On Examination line, and those who are found to be grossly unstable are given intensive psychological and psychiatric study, in order to make the best possible disposition of their cases. Since August 1944, 41,000 Sign-On screenings have been conducted at the New York Port alone, and from these interviews it is hoped that we can eventually determine the various personality patterns of successful seamen, and, on the basis of these find-

ings, reorient our selection program so as to obtain for the industry the most suitable and most effective types of personality.

About the eventual picture of the successful seaman that may emerge, we can now only hazard a guess, but sufficient clinical evidence has already been accumulated to indicate that the negative traits, traditionally attributed to peace-time merchant seamen, do not hold true for the present, war-time cross-section of the industry. On the whole, the merchant mariners of the moment are capable, industrious men, who compare favorably with any other industrial group doing work requiring the same degree of capacity and skill.

DISCUSSION OF THE PAPERS

Dr. Rose G. Anderson (*Psychological Service Center, New York*): I should like to preface any comment on these significant papers by indicating my position with respect to the use and interpretation of personality inventories.

Such personality tests have three functions (1) screening; (2) guiding the clinical interview by high-lighting areas needing further investigation, or eliminating certain areas from further consideration; and (3) providing supplementary material in arriving at clinical judgments Such tests are never a substitute for the individual interview nor for clinical judgment.

For one reason, the items as rated cannot be accepted as objective fact, but must be evaluated in terms of the determinants of each individual's response. The response may be one of complete naïveté, *i.e.*, uncritical marking of the items because of lack of appreciation of their implications. It may be that of the conscious malingerer who, especially in military service, has a motive for presenting himself as unqualified for specific responsibilities. In other cases, especially in connection with employment, the conscious motive to answer in the supposed favorable direction influences the rating of the items and distorts the picture in various ways.

Further, we have that considerable proportion of individuals who lack self-insight in varying degrees, whose ratings are weighted according to the direction and degree of their errors of self-estimate.

The optimum value of the personality inventory results from its use as an adjunct to the interview, after *rapport* has been established with a consultant from whom the individual desires insight and guidance, and in whom he feels sufficient confidence to reveal himself unreservedly.

In view of this position, I raise the question whether any personality test should attempt to diagnose the wide range of mental deviations included in the Minnesota Multiphasic test. It seems comparable to the common lay concept of aptitude testing We frequently have individuals requesting *an* aptitude test. We explain that identification of aptitudes involves a range of aptitude tests. Also, that the selection of the appropriate tests for any specific individual depends upon information about his experience, training, etc.

To attempt to diagnose many mental deviations by one measure seems analogous to attempting to measure many aptitudes by one test, with the exception that, in the latter case, we are looking for positive evidence, rather than negative.

In the Minnesota Multiphasic Inventory, much irrelevant material is included for any specific case. This is apparent from inspection and from the mutually exclusive diagnostic categories. However, I should like Major Abramson's comment on whether he has not found the test more useful in guiding and short-cutting the interview than in diagnosis It would appear from his data that an appreciable number of items could be eliminated from the scale, since the distinctive pattern of the individual profile persisted in the limited number of items answered consistently in the repetition of the test.

According to the reports in the literature, there are a number of high positive correlations between diagnostic categories, *e.g.*, as high as + .55 between hysteria and depression in normal subjects, and as high as + .71 between hysteria and hypochondriasis in hospital patients. The evidence points to the possibility of both shortening the scales and differentiating the categories, by eliminating less discriminatory items.

This would seem especially desirable when the scale is used for screening. That a shorter number of critical items results in satisfactory screening is demonstrated by Dr. Killinger's report.

Mr. Arthur E. Traxler (*Educational Records Bureau, New York*): My comments represent the viewpoint of one interested in the guidance of normal young people. My main interest is in those aspects of personality which can be stated in the every-day terminology of teachers, counselors, and school psychologists. As is true of many other persons in this field, I tend to take a conservative view of the value of personality testing, as far as the application of these techniques outside clinical situations is concerned. In the first place, we have the perennial question concerning whether or not there are generalized personality traits and whether by means of personality tests we can sample anything that is stable under varying conditions

Even if certain aspects of personality are stable, we still find personality tests of rather limited value, because all our measuring techniques in this field are more or less experimental. Projective techniques are so nebulous, and the interpretations are so involved and technical, that their successful use seems to be beyond anyone except the expert. The basic method of most non-projective techniques (open to all the well-known limitations of the questionnaire method) may be helpful to a clinician, while of much less value in other hands

It is obvious, of course, that the term, "non-projective," covers a wide variety of specific techniques, some of which have little in common. It includes instruments the purposes of which are well disguised, those whose purposes are partially disguised, and those whose purposes are not hidden at all from a moderately intelligent subject. In general, the attempts to measure personality by means of non-projective techniques whose purposes are not apparent to the subject have not been fruitful. So, during the last fifteen years, the majority of those constructing personality tests have based their techniques upon the self-inventory, or psychoneurotic questionnaire, in which the individual is required to respond with "yes," "no," or "doubtful," to a series of questions concerning how he acts or feels in a variety of situations. Some of these questions have undoubtedly been subjected to very careful scrutiny, for they were first used by Woodworth, about 1920, and the same questions have appeared in many different inventories since that date

The application of the multiple-scoring technique has greatly increased the kinds of information one can obtain through the administration of a single series of questions to a subject. Many different scales have been set up and standardized on the basis of this type of atomistic approach to the study of personality, and some of these tests have been developed through admirable procedures of test construction. Nevertheless, the validity of all these measures is almost wholly dependent upon *rapport;* upon the honesty and fairness of the individual taking the test; and his ability to appraise accurately his own reactions and feelings.

The Minnesota Multiphasic Personality Inventory is one of the newest of the personality questionnaires set up and standardized for clinical use. The first published data on the inventory appeared in the literature less than five years ago, and work on it is still being done. A considerable number of the questions in this inventory have appeared in other tests, but it includes more questions and a wider variety of questions than most of the other personality tests. The original technique of individual administration through the use of cards may be an improvement over the usual procedure, when the subjects are clinical cases. A group booklet has recently been prepared, and the test has been adapted for machine-scoring.

Hathaway and McKinley have published four articles on the inventory. The first outlined the general nature of the measurement project in which they were engaged, and the other three described the development, respectively, of the scales

for hypochondriasis, symptomatic depression, and psychasthenia. As far as I can tell from careful reading of the articles, the procedures they used involved good, defensible techniques of test construction. The main weakness seemed to be in the small size of the criterion group. fifty for each of the first two scales, and only twenty for the third one. The care with which the groups were chosen probably makes the small size of the groups less of a limitation than it would otherwise be.

Hathaway and McKinley's description of the procedure employed in choosing the criterion groups seems to imply that there is a dichotomy between normal and abnormal subjects with respect to such a concept as psychasthenia. I would be inclined to question that assumption, if it is actually made by them The distribution of individuals with regard to any of these aspects of personality is, I believe, a continuum Their criterion cases were not different in *kind* from the normal subjects They simply represented extreme deviations from the mean of the normal population, even though they were clinical cases.

Six additional scales for the inventory have been published, although, apparently, detailed information on the procedures used in constructing these scales has not been made available

One of the points made by Major Abramson was that the shape of the profile of scores on the Minnesota Multiphasic Personality Inventory is more important than the actual value of the scores I think that this statement could be extended to include all personality inventories yielding multiple scores

There was an impressive amount of agreement between the profiles obtained from several of Major Abramson's cases before and after the administration of alcohol In view of the reliability coefficients elsewhere reported, it appears that the agreement is as close as one would expect to obtain from two administrations of the multiphasic test to the same individuals, without the use of alcohol in the interim

Major Abramson stated that the multiphasic test provides an excellent shortcut to obtain a psychosomatic history. This is, I believe, one of the chief values of personality inventories in general

Dr Killinger made the statement that, in the first personality inventory devised for screening procedures in the War Shipping Administration, the wording of some of the items was too difficult, and that it was found advisable to revise the inventory. The wording of all personality inventories probably needs careful study. Some of these inventories may tend to be disguised intelligence tests for certain individuals.

I would like to enter a plea for more research on specific measurement devices of the type represented in these papers. One is constantly amazed by the extensive use of hundreds of measuring devices, mental, achievement, aptitude, interest, personality, which have been the subject of almost no research except that carried on by the authors at the time they constructed the tests. By and large, these instruments are taken on faith. The field is so nebulous, and the relationship between personality-measuring devices and their avowed purposes is so subtle, that one cannot validate these instruments by a common-sense process of inspection. It is necessary to study the scores in relation to expert judgment and long-time case records, before the worth of the tests can be appraised.

Although the Minnesota Multiphasic Personality Inventory has thus far been a clinical instrument, it is to be hoped that eventually its use may become somewhat broader, and that it will be feasible to recommend it for experimental use in guidance situations in educational institutions. It is research such as that here reported which will decide for us the worth of this instrument and other personality inventories. It would be very helpful, for example, to have additional information on the intercorrelation of the scales for the multiphasic test; data on the correlation of the scores on these scales with intelligence test results; and on other measures.

NON-PROJECTIVE PERSONALITY TESTS

PART II

THE CORNELL INDICES AND THE CORNELL WORD FORM:

THE CORNELL INDICES AND THE CORNELL WORD FORM:

1. CONSTRUCTION AND STANDARDIZATION*

By Bela Mittelmann and Keeve Brodman

The Cornell Service Index,[1, 2] the Cornell Selectee Index,[3, 4, 5] and the Cornell Word Form were designed: (a) to differentiate, quantitatively, individuals with personality and psychosomatic disturbances from the rest of the population; and (b), especially in the case of the Indices, to facilitate qualitative diagnosis of most of these disorders by the completeness of the information about the subject's symptoms.

The Cornell Indices ascertain very directly, by asking questions, whether the subject does, or does not, claim to have specific symptoms. Questions are of the type asked in psychiatric interviews. The Cornell Word Form aims at detecting the presence of personality and psychosomatic disturbances by an indirect method. It is useful in situations where strong motivation may make responses to direct questions unreliable. In all tests, about half the items refer to psychosomatic symptoms, and the rest to behavioral and emotional disturbances.

All of these tests are self-administered, and may be given to any number of individuals simultaneously. Literacy is the only intellectual necessity for completing and scoring. Either may be completed in ten minutes and scored within one minute.

All items and scoring methods were standardized on population samples, of from two to four hundred subjects, collected in various sections of the country. These samples were evaluated separately, procedures being decided upon only when they were applicable to all samples, even if others yielded better results on single samples.

The tests were validated on subjects grouped according to the following criteria: (1) acceptance or rejection after psychiatric interview at induction; (2) classification as "normal," or "severely psychoneurotic," on interview during military service; (3) performance in the

* This and the following two papers were prepared at the New York Hospital and the Departments of Medicine (Neurology) and Psychiatry, Cornell University Medical College; and the Psychiatric Division, Bellevue Hospital, New York, N. Y., with the technical assistance of Margaret Meixner.

The work described in these papers was done under a contract, recommended by the Committee on Medical Research, between the Office of Scientific Research and Development and Cornell University.

This study was aided by a grant from the Josiah Macy, Jr., Foundation.

armed forces. The latter group included students at Officer Candidate Schools and military personnel who had been, or were about to be, discharged because of neuropsychiatric disability.

Direct Tests

There are two Cornell Indices. One is the Cornell Selectee Index, containing questions applicable to men and women between the ages of eighteen and forty; the other is the Cornell Service Index, containing additional items applicable only to men who have completed at least six weeks of military training.

The wording of items is significant. Some refer, in a generalized manner, to a somewhat vague complaint, such as, "Do you suffer from stomach trouble?" Others aim at more specific information, such as, "Do you frequently suffer from nausea (sick to your stomach)?" Other questions concerning the same symptom complexes were couched in mild or severe language; *e.g.*, "Are you considered a nervous person?", and "Does every little thing get on your nerves and wear you out?"

Three principles were used in selecting items to be incorporated into the tests. (1) Significant differentiation between those with neuropsychiatric disturbances, and ostensibly healthy persons. The majority of the items yielded significant results, whether *chi* square or critical ratio was used. For example, the question, "Do you suffer badly from frequent severe headaches?", had a *chi* square value of 39.06 and critical ratio of 12.1. (2) Questions concerning specific defects, such as fits or convulsions, even if they occurred infrequently among the neuropsychiatrically unfit, were included, if they did not segregate a large number of "false positives"* in any of the samples. These are known as "stop questions." (3) If two questions concerning the same symptom are each answered "Yes," by, for instance, two per cent. of the normals and ten per cent. of the neuropsychiatrically unfit group, and the ten per cent. constitutes the same persons in both cases, while the two per cent. constitutes different persons answering "Yes" to each of the two questions, only one of the questions is used. On the other hand, if the ten per cent. in each case mainly constitutes different people, and the two per cent. represents mainly the same persons answering "Yes" to both questions, then both questions are used. This principle is also applied to the relation between the total score and the incidence of a "stop question." If subjects in the neuropsychiatrically unfit group answer "Yes" to a given question, but are nearly always detected on

* A healthy individual who is indicated as being abnormal by the test is called a "false positive."

the basis of total score, the item is used as a "stop question" only if the increase in false positives is not over one per cent.

The questions of the Cornell Service Index are grouped according to symptom complexes, so as to facilitate clinical interpretation. The least disturbing questions are placed at the beginning of each group, and the least disturbing groups at the beginning of the form. The order in which the groups appear in the "Index" is given in TABLE 1.

TABLE 1

a. Questions 1–3 are introductory and neutral.
b. Questions 4–17 concern defects in adjustment to military groups, expressed as feelings of fear and inadequacy.
c. Questions 18–20 and 24–30 concern pathological mood reactions, especially anxiety and depression.
d. Questions 21–23 concern neurocirculatory psychosomatic symptoms.
e. Questions 31–37 concern pathological startle reactions.
f. Questions 38–49 concern a variety of other psychosomatic symptoms.
g. Questions 50–63 concern hypochondriasis and asthenia.
h. Questions 64–74 concern gastrointestinal psychosomatic symptoms
i. Questions 75–79 concern excessive sensitivity and pathological suspiciousness.
j. Questions 80–92 concern symptoms of troublesome psychopathy.

An item analysis of 352 selectees, accepted for service in the armed forces after neuropsychiatric interview, was compared to the item analysis of 210 soldiers discharged for neuropsychiatric disability. TABLE 2 shows the distribution of critical ratio values for all the sixty-four

TABLE 2

DISTRIBUTION OF CRITICAL RATIO VALUES* OF THE 64 ITEMS OF FORM N FOR 210 SOLDIERS DISCHARGED FOR NEURO-PSYCHIATRIC REASONS COMPARED TO 352 SELECTEES ACCEPTED AFTER BRIEF PSYCHIATRIC INTERVIEW

Critical Ratio Values	Number of Items
2.5– 2.9	2
3.0– 3.9	5
4.0– 4.9	9
5.0– 5.9	6
6.0– 6.9	4
7.0– 7.9	6
8.0– 8.9	7
9.0– 9.9	8
10.0–10.9	3
11.0–11.9	8
12.0–12.9	3
13.0–13.9	2
14.0–14.9	1

* Critical ratio values were determined by the Fulcher-Zubin Method.[8]

items. It will be seen that two items have critical ratios between 2.5 and 2.9, while the remaining sixty-two have critical ratios of 3.0 or higher. TABLE 3 shows the incidence of important symptoms of psychiatric significance. The general selectee population represented by the 1000 registrants is compared to the 210 soldiers discharged by the Army. These two groups are compared on each of the "stop questions."

TABLE 3

INCIDENCE OF EACH 'STOP QUESTION' AMONG THE GENERAL SELECTEE POPULATION AND THOSE DISCHARGED BY THE ARMY BECAUSE OF NEUROPSYCHIATRIC DISABILITY

'Stop Question'	General Selectee Population (N–1000) %	Discharged by Army Because of Neuropsychiatric Disability (N 210) %
5. Have you ever gotten into serious trouble or lost your job because of drinking?	0.7	10 9
15. Have you ever had a fit or a convulsion?	1.8	19.0
40. Are you a bed-wetter?	3.0	7.1
45. Were you ever a patient at a mental hospital?	0 7	13 8
50. Are you a sleep-walker?	1.0	5 2
55. Do you suffer badly from frequent loose bowel movements?	2.1	10.5
59. Did you ever have a nervous breakdown?	0 7	38 6
60. Has any doctor ever told you that you had ulcers of the stomach?	0 4	11 4
62. Do you drink more than 2 quarts of whiskey a week?	0 4	5 7
63. Have you been arrested more than 3 times?	0.5	5.2

Indirect Methods

The indirect tests aim at detecting personality and psychosomatic disturbances, while preventing the subject's becoming aware of this purpose. Three such methods were used experimentally. In one, the subject expressed a preference for certain occupations His choices were scored according to masculinity or femininity, the result to indicate sexual disturbances. Another aimed at measuring decisiveness and self-confidence by having the subject rate himself as to his ability to handle various situations. These two tests yielded useful qualitative information, but were limited in the significance of their quantitative differentiation, and, therefore, at present, are not being used.

The third method, the Cornell Word Form, has both quantitative and qualitative usefulness. The subject is presented with a list of

stimulus words, next to each of which are two other words. He is asked to choose the one he thinks goes better with the stimulus word. Some of the choices are comparatively obvious in their implication, for example, SLEEP—comfort, restless. In others, it is not obvious; for example, MOTHER—mine, woman. Words of the latter type are in the majority. Both yield significant differentiating value, although most of the items with the highest critical ratios belong to the obvious type. However, in the total number of items, their direction may be overlooked, and the subject may give informative responses, in spite of a desire to falsify.

In general, the differentiating value of the items of this test was not as high as that of the direct test, but on the basis of total score, significant quantitative differentiation can be made between normals and those with neuropsychiatric disorders.

Principles of Scoring

The following principles were used in determining scoring methods. (1) A scoring level was determined at which the maximum number of the "moderately severely," and "severely" psychoneurotic persons was detected, without also segregating an excessive number of "false positives." (2) A scoring level at which the number of false positives was minimal, yet at which the majority of "severely" psychoneurotic persons was detected. (3) In the case of the Indices, additional segregations of those with specific complaints (*e.g.*, epilepsy), by the use of "stop questions" as scoring criteria.

THE CORNELL INDICES AND THE CORNELL WORD FORM:

2. RESULTS

By Arthur Weider and David Wechsler

The Cornell Selectee Index

The purpose of the Cornell Selectee Index was to furnish a quick, reliable means of detecting, at induction, individuals who are likely to develop psychoneuroses and psychosomatic disturbances.

The Cornell Selectee Index was first given to 1000 persons falling into the following four categories: those accepted by neuropsychiatric interview at induction in New York (450) and in Boston (450); those rejected after neuropsychiatric interview at induction in New York (50) and in Boston (50).

Table 4 presents the three correlations determining reliability for three cut-off points The 1000 cases gathered at induction stations were used for these correlations. The table reveals that all tetrachoric correlations[7] are significantly high, indicating the reliability of the Cornell Selectee Index. The usual technique of splitting a questionnaire into two forms (odd-even) was the method used in ascertaining reliability measures.

Table 4

Tetrachoric Reliability Coefficients*

General Selectee Population
(N-1000)

	Cut-off used
0.83	3
0.93	7
0.93	9

* Correlations were computed by using the split-half (odd-even) technique for reliability

Table 5 presents a four-fold contingency table, the results of the Cornell Selectee Index, as against the decision of the induction neuropsychiatrist. It will be seen that, of 100 selectees rejected and 900 accepted for military service after a brief psychiatric interview, the Cor-

nell Selectee Index score agreed with the neuropsychiatric opinion in rejecting eighty-nine selectees and accepting 790. The Index missed eleven men rejected after interview, while 110 men are considered "false positives." The tetrachoric correlation is .89, indicating a high degree of relationship between the two methods. Eighty-nine per cent. of those rejected after neuropsychiatric interview were detected on the basis of Index score, while twelve per cent. of ostensibly healthy persons were also segregated.

TABLE 5

FOUR-FOLD CONTINGENCY TABLE SHOWING THE NUMBER OF SELECTEES ACCEPTED AND THE NUMBER OF SELECTEES REJECTED BY BRIEF PSYCHIATRIC INTERVIEW WHO ARE IDENTIFIED BY FORM N

	Pass	Fail*	Total
Rejected	Test 11 Miss	Both Agree 89 To Reject	Interview 100 Rejected
Accepted	Both Agree 790 To Accept	Test 110 False Positive	Interview 900 Accepted
Total	Test 801 Pass	Test 199 Fail	Total 1000 Cases

Tetrachoric Correlation (r_t = 0.89)
Rejects Detected 89%—False Positives 12%

* A subject is considered as failing Form N if a score of 15 or one or more "stop questions," is achieved.

TABLE 6 is another contingency table which shows the performance of the Cornell Selectee Index for a group of 204 men discharged from the service because of neuropsychiatric disability, and for a group of 406 men accepted for military training. The tetrachoric correlation is 0.93, indicating a high degree of relationship between the two methods Eighty-two per cent. of those discharged were detected on the basis of Index score, while seven per cent. of ostensibly healthy persons were also segregated.

Two scoring methods can be applied to the test, depending upon the demands of the situation in which it is used. Method A, using fifteen significant answers, or one or more "stop questions" as criteria, detects persons suffering from neuropsychiatric and psychosomatic disorders, even when they are not severe enough to make military service impossible. A small number of healthy persons is also included in this group, because of misinterpretation of test items. Method B uses twenty-five significant answers, regardless of the number of "stop ques-

TABLE 6

FOUR-FOLD CONTINGENCY TABLE SHOWING THE NUMBER OF MEN DISCHARGED FROM THE SERVICES BECAUSE OF NEUROPSYCHIATRIC DISABILITY AND THE NUMBER OF MEN ACCEPTED FOR MILITARY TRAINING, IDENTIFIED BY FORM N

	Pass	Fail*	Total
Discharges	Test 31 Miss	Both Agree 173 To Discharge	Interview 204 Discharge
'Normals'	Both Agree 379 To Accept	Test 27 False Positive	Interview 406 Accepted
Total	Test 410 Pass	Test 200 Fail	Total 610 Cases

Tetrachoric Correlation (r_t – 0 93)

Discharges Detected 82%—False Positives 6%

* A subject with a score of 15, or one or more "Stop Questions" is considered as failing Form N

tions" as the scoring criterion. It screens only those individuals suffering from severe neuropsychiatric and psychosomatic disturbances. Many persons who are ill will not be detected, but those detected are ill.

The Index was given to another group of 1000 selectees at induction stations in various parts of the country.

TABLE 7 shows the percentage of induction populations earmarked and detected. Data for the two above, and a third scoring level, are given. It will be seen that, by using scoring method A, sixty to eighty-eight per cent. of neuropsychiatrically unfit selectees are detected, while

TABLE 7

PER CENT OF INDUCTION POPULATIONS EARMARKED AND DETECTED AT THREE SCORING LEVELS OF THE CORNELL SELECTEE INDEX

Group	Area	Scoring Levels 15–1 Stop %	15 %	25 %
100 Psychiatric Accepts	North East	9	5	0
100 Psychiatric Rejects	North East	84	78	42
100 Psychiatric Accepts	East	17	8	2
50 Psychiatric Rejects	East	66	52	20
100 Psychiatric Accepts	East	15	8	1
100 Psychiatric Rejects	East	60	45	24
100 Psychiatric Accepts	Mid West	10	6	0
50 Psychiatric Rejects	Mid West	88	76	38
200 Psychiatric Accepts	North West	19	7	0.5
100 Psychiatric Accepts	North West	63	50	16

between nine and nineteen per cent. of ostensibly healthy persons are also segregated. Similarly, with scoring method B, sixteen to forty-two per cent. of the neuropsychiatrically unfit, and up to two per cent. of ostensibly healthy persons, are segregated.

TABLE 8

PERCENTAGE OF NEUROPSYCHIATRICALLY UNFIT (AS ASCERTAINED BY INTERVIEW METHOD AND MILITARY PERFORMANCE) "SCREENED" BY THE SELECTEE INDEX

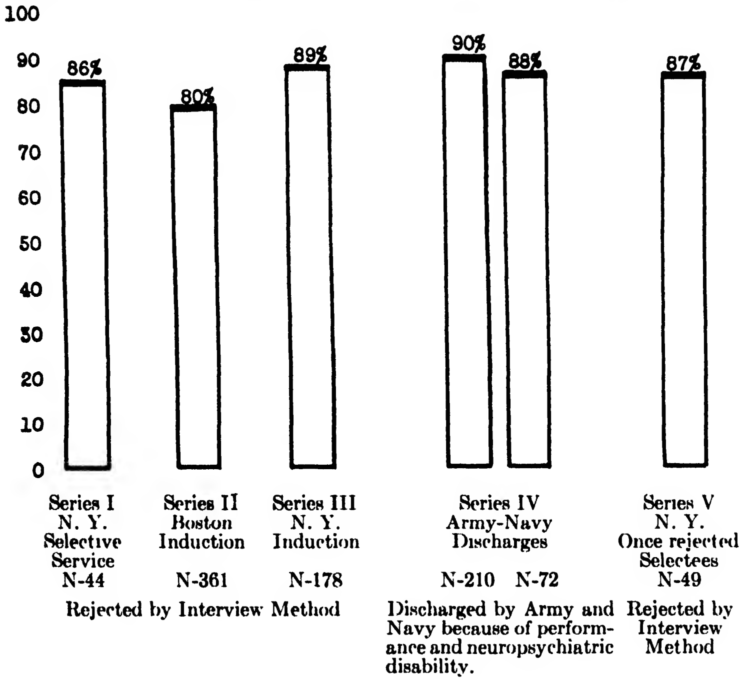


TABLE 8 shows the percentage of neuropsychiatrically unfit (as ascertained by interview or military performance) screened by the Index. It will be seen that eighty to ninety per cent. are detected, while not more than twenty-five per cent. of ostensibly healthy persons are earmarked.

In TABLE 9, is shown the percentage screened of those in the previous table with various neuropsychiatric and psychosomatic disturbances, as so diagnosed by the induction neuropsychiatrist or medical officer.

As the need for a similar instrument for use with military personnel became apparent, the Index was adapted to meet the new situation.

TABLE 9

PERCENTAGE OF NEUROPSYCHIATRICALLY UNFIT (AS ASCERTAINED BY INTERVIEW METHOD AND MILITARY PERFORMANCE) "SCREENED" BY THE SELECTEE INDEX

Psychiatric and Psychosomatic Disturbances*	Per cent. "Screened"
Psychoneurosis—	
Mixed	82
Anxiety	71
Miscellaneous	87
Constitutional Psychopathic States	82
Psychotic States—	
Dementia Praecox	74
Miscellaneous	91
Ulcer	96
Migraine	92
Asthma	93
Arterial Hypertension	53
Enuresis	94
"Immaturity"	77
Stuttering and Stammering	75
Epilepsy and Convulsive Disorders	100
Miscellaneous	93

* Classifications as quoted by the various Induction Station and military hospitals

The Cornell Service Index

The Cornell Service Index was designed to give to the armed forces a quick, reliable method for obtaining important facts of the neuropsychiatric history, with a simple method of scoring these items, which would differentiate individuals with serious personality problems from the rest of the population.

Results for the following groups will be presented:

1. 539 men in military service judged to be "normal," on the basis of psychiatric interview or military performance.
2. 142 severely psychoneurotic men discharged from the services.
3. 260 moderately severely psychoneurotic men placed, or about to be placed, on Limited or Special Assignment.
4. 39 mildly psychoneurotic men.

Three scoring methods were developed for use in various situations. Method A, using an Index score of twenty-three or more, screens the majority of those suffering from serious neuropsychiatric and psychosomatic disturbances. Method B, using an Index score of 13 or more, screens almost all those with serious, and the majority of those with mild, neuropsychiatric and psychosomatic disturbances; a few ostensi-

bly healthy persons are screened. Method C, using an Index score of thirteen or more, and the occurrence of one or more "stop questions" as scoring criteria, screens, in addition, individuals who claim to have especially significant disorders, *e.g.*, fits. Affirmative responses to these "stop questions" are useful in pointing out initial leads in the interview.

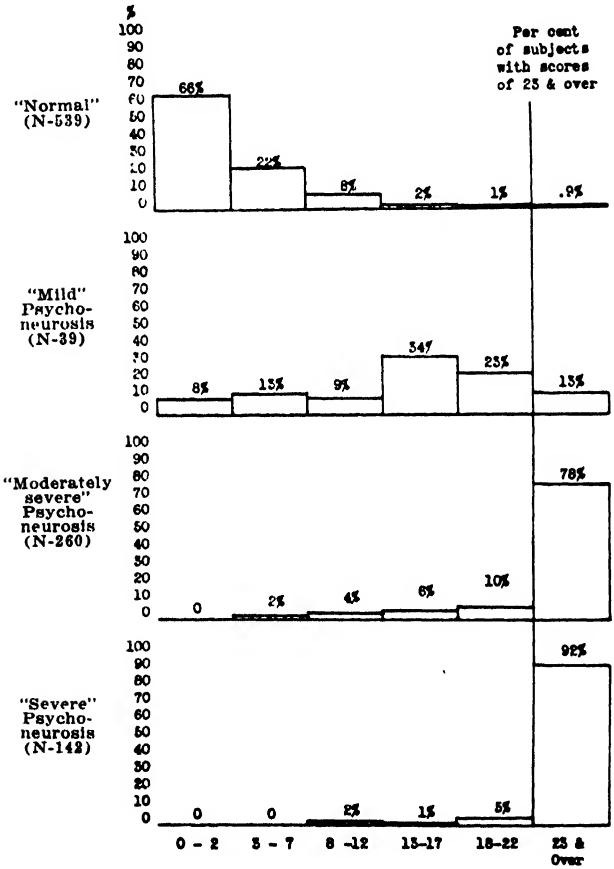


FIGURE 1
"INDEX" SCORES OF SELECTED POPULATIONS

FIGURE 1. The men were divided, according to score, into three groups: those with scores of twelve or less; those with scores of from thirteen to twenty-two; and those with scores of twenty-three and above. The low score group contains ninety-six per cent. of the "normals"; the middle range, fifty-seven per cent. of the mildly psychoneu-

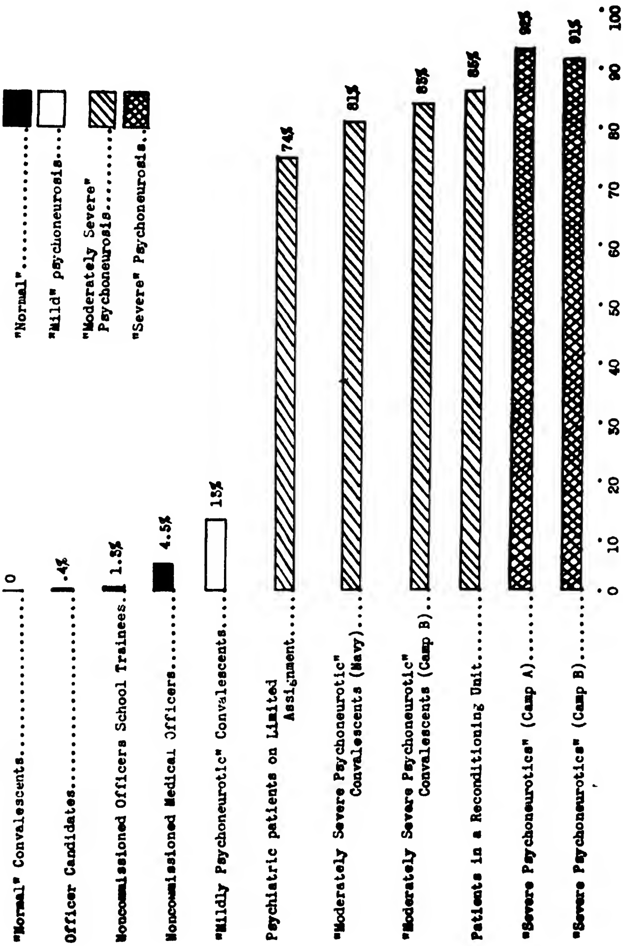


FIGURE 2

PER CENT. OF SELECTED POPULATIONS WITH SCORES OF 23 AND OVER

rotic; and the high score range, ninety-two per cent. of the severely psychoneurotic. There is, therefore, on the basis of the Index score, a well-defined difference between the normal, and severely psychoneu-

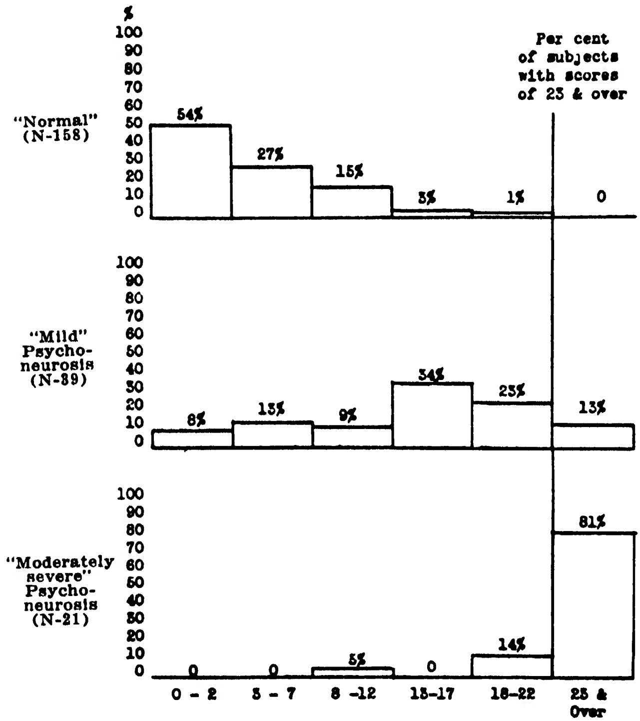


FIGURE 3
"INDEX" SCORES OF CONVALESCENT POPULATIONS

rotic, subjects. The moderately severely psychoneurotic obtain scores similar to those of the severe.

FIGURE 2. The Index was given to convalescent patients (men of military age), in several general hospitals. Eighty-one per cent. of the moderately severe group are detected at a scoring level of twenty-three, while none of the normals falls into this group. At a scoring

level of thirteen, ninety-five per cent. of the deviant group, four per cent. of ostensibly healthy persons, and seventy per cent. of mildly psychoneurotic persons, are segregated.

FIGURE 3 shows the percentage of the selected populations which was detected at a scoring level of twenty-three.

While falsification of responses in these questionnaires is rare, it was felt that a method, the results of which would not become unreliable in cases of strong motivation, would be of value.

The Cornell Word Form

The Cornell Word Form was devised to effect the differentiation of "normals" from those with neuropsychiatric and psychosomatic disorders, in a manner not apparent to the subject.

TABLE 10

PERCENTAGE OF VARIOUS POPULATIONS EARMARKED AND DETECTED AT THREE SCORING LEVELS OF THE CORNELL WORD FORM

Scoring Level	Navy Population used in Item Analysis		Mid-West Induction Station		Western Induction Station	
	Psychoneurotic N-100	Normal N-100	Reject N-50	Accept N-100	Reject N-100	Accept N-200
	%	%	%	%	%	%
Method A—9	51	2	46	5	15	4
Method B—5	78	15	78	25	47	10
Method C—5-2*	85	15	82	25	52	10

* Irregular responses are those in which neither or both words are indicated

There are three scoring methods. Method A uses a stencil with which the number of significant responses is counted. A scoring level of nine, or more, screens between thirteen and fifty per cent. of persons with neuropsychiatric disturbances and two to five per cent. of ostensibly healthy persons. Method B uses a scoring level of five, or more, detecting forty-seven to eighty per cent. of those with neuropsychiatric disturbances and ten to twenty-five per cent. of ostensibly healthy persons. Method C uses a scoring level of five and the occurrence of two or more irregular responses; that is, where both, or neither, of the words are chosen. Thus, fifty-two to eighty-six per cent. of those with neuropsychiatric disturbances are screened with no increase in false positives. This method may be modified to use a score of five, plus one irregular response, as scoring criteria, in which case the detection rate will be raised by five per cent. with an additional one per cent. false positives. TABLE 10 indicates the efficiency of the test in various samples.

THE CORNELL INDICES AND THE CORNELL WORD FORM:

3. APPLICATION

By Harold G. Wolff

The Cornell Indices and the Cornell Word Choice Form represent an important development in test construction. They are based on clinical experience, psychiatric principles and psychological statistical techniques. They differentiate significantly between "normal" and neuropsychiatrically unfit persons. Furthermore, these tests are easily and rapidly administered and scored and may be given to any number of individuals simultaneously.

The Indices are useful in situations in which it is necessary to collect quickly items of neuropsychiatric and psychosomatic data. This is especially true when it is important to discover those who, among a large number of individuals, require psychiatric appraisal and care, when sufficient time is not available for lengthy psychiatric interview. The Cornell Word Choice Form may be used in addition to, or instead of, the Indices, if an indirect procedure is more suitable to the situation.

The Indices are applicable: (1) When it is desirable for the psychiatrist to collect a significant volume of neuropsychiatric data in a minimum of time. (2) When it is necessary to collect items of neuropsychiatric data, and psychiatric interview is not possible. In this situation, the data can be collected on the Indices for future evaluation by the neuropsychiatrist. (3) When it is necessary to segregate by statistical methods individuals with neuropsychiatric and psychosomatic disturbances, as in large scale surveys.

Specifically, the Indices may be applied in military and civilian situations, as listed below. These applications were found useful after evaluation of approximately 15,000 cases.

MILITARY:

A. *Induction Stations.* The Cornell Selectee Index is administered at induction stations for the purpose of collecting items of neuropsychiatric history useful in diagnosis and disposition. Both the total score and responses to single items are used in establishing diagnosis.

The Index places at the disposal of the neuropsychiatrist a volume of neuropsychiatric data, adequate for the rapid appraisal of a large number of selectees in a short space of time.

B. *Neuropsychiatric Wards and Clinics of Military Hospitals.* The Cornell Service Index enables the neuropsychiatrist to obtain significant items of neuropsychiatric history, rapidly and without any expenditure of his time and energy. An enlisted man may be assigned to the task of administering the Index before the patient is seen in interview. The neuropsychiatrist then has available at the beginning of his interview pertinent data to help him in determining diagnosis and prognosis.

C. *Medical and Surgical Wards of Military Hospitals.* Many patients with structural disorders also have emotional disturbances which interfere with rapid convalescence. These disturbances are sometimes not obvious and may remain unrecognized as factors contributing to prolonged convalescence. Medical and surgical officers do not have enough time for thorough investigation of the personalities of their patients, nor do neuropsychiatrists have time to see all medical and surgical patients in consultation. The Cornell Service Index can be used as a quick and reliable means of exposing emotional disorders in medical and surgical patients,[8] preliminary to referral for neuropsychiatric consultation.

CIVILIAN:

A. *Neuropsychiatric Wards and Clinics.* Psychiatric clinics often have referred to them larger numbers of patients than can be cared for adequately by the available staff. One phase of psychiatric investigation that is time-consuming is the accumulation of a sufficient body of historical data and information about present complaints. This material can be collected quickly by means of the Cornell Selectee Index, and placed at the disposal of the neuropsychiatrist, with no expenditure of his time. He can then attack therapeutic problems more quickly.

B. *Medical and Surgical Wards and Clinics.* The use of the Cornell Selectee Index in civilian hospitals[9] parallels that of the Cornell Service Index in military hospitals.

C. *Industry.* In industry, the Cornell Selectee Index can be of use to the physicians of the medical department in determining which employees need guidance for emotional disturbances or lowered morale. The Index is not recommended as a criterion by which an employee is hired or discharged; although, in certain situations and for certain purposes, it may be applied in placing employees suitably.

D. *Returning Veterans*. The Cornell Service Index can be useful in the rapid neuropsychiatric evaluation of returning servicemen, preliminary to proper placement in civilian jobs. Because the Index is self-administered and quickly evaluated, it is an aid in the rapid processing of large numbers of veterans.

E. *Research*. Until the present time, adequate analysis of large numbers of individual psychiatric case histories has not been practicable, because of the length of time involved in acquiring data. Even when data are gathered by neuropsychiatrists working in teams, it has been difficult to compare cases, because the body of information has not been standardized. The Cornell Indices offer a simple and expeditious method for obtaining adequate standardized psychiatric data which may be subjected to statistical analysis.

REFERENCES

1. **Weider, A., K. Brodman, B. Mittelmann, D. Wechsler, & H. G. Wolff**
   1945. (April). The Cornell Service Index: A Method for Quickly Assaying Personality and Psychosomatic Disturbances in Men in the Armed Forces. War Medicine **7**: 209.
2. **Warner, Lieut. N. (MC, USNR), & M. W. Gallico, Lt. (j. g.) (USNR)**
   1945. (April). Use of the Cornell Service Index in the Evaluation of Psychiatric Problems in a Naval Hospital. War Medicine **7**: 214.
3. **Weider, A., B. Mittelmann, D. Wechsler, & H. G. Wolff**
   1944. (Jan. 22). The Cornell Selectee Index: A Method for Quick Testing of Selectees for the Armed Forces. J. A. M. A. **124**: 224.
4. **Weider, A., B. Mittelmann, D. Wechsler, & H. G. Wolff**
   1944. (Jan.). The Cornell Selectee Index: Short Form to be Used at Induction, at Reception and During Hospitalization. Proceedings, Second Brief Psychotherapy Council.
5. **Harris, Lt. Comdr. H. J., (MC, USNR)**
   1945. (May). Functions of a Psychiatrist in a Navy Yard. United States Naval Medical Bulletin **44**: 1036.
6. **Fulcher, J. S., & J. Zubin**
   1942. The Item Analyzer: A Mechanical Device for Treating the Four-Fold Table in Large Samples. J. Applied Psychology **26**: 511.
7. **Chesire, L., M. Saffir, & L. L. Thurstone**
   1933. Computing Diagrams for the Tetrachoric Correlation Coefficient. University of Chicago Press, Chicago.
8. **Warner, Lieut. N. (MC, USNR), & M. W. Gallico, Lt. (j. g.) (USNR)**
   The Occurrence of Psychoneurotic Symptoms on the Various Services of a Naval Hospital. United States Naval Medical Bulletin. In press.
9. **Mittelmann, B., A. Weider, K. Brodman, D. Wechsler, & H. G. Wolff**
   1944. Personality and Psychosomatic Disturbances in Patients on Medical and Surgical Wards: A Survey of 450 Admissions. Proceedings, Association for Research in Nervous and Mental Disease. Williams & Wilkins, Co. Baltimore.

THE CORNELL SELECTEE INDEX*

AN AID IN PSYCHIATRIC DIAGNOSIS

By Harold J. Harris

New York, N. Y.

In February 1944, the Cornell Selectee Index, both forms C and N, were made available to me through the courtesy of the authors. The work being done consisted of psychiatric examinations and reexaminations of certain designated civilians, and consultations for personnel of the Navy, including the Marine Corps and Coast Guard, at a Navy Yard. The volume of this work was so great as to make essential the use of time-saving methods. There are about 75,000 civilian employees in the Yard; perhaps 1,000 Marines stationed in the Yard, most of whom have had combat duty; about 2,000 officers and enlisted men of the Navy attached to the Yard; and from 10,000 to 50,000 officers and men attached to ships in the Yard. Thus, the total population served by one psychiatrist is seen to vary from about 85,000 to 125,000. The total number of psychiatric examinations made each month has ranged from approximately 600 to well over 800. The largest number in any one day has been 69. The daily average has been about 30 for the past six months or more. More than 9,000 persons have executed the forms prior to psychiatric examination in the past twelve months. It is obvious that these diagnostic interviews must be done rapidly. Even with the use of the Index, from 15 to 45 minutes may have to be devoted to some important consultations. This necessitates allotting but a very few minutes to the average, less important, interview.

Without the information furnished by the use of the Cornell Selectee Index, any degree of accuracy in such a large number of examinations would be virtually unattainable, for it serves as a starting-point from which additional information is quickly added. The questions listed are those likely to be asked in any psychiatric interview.

The method of use has been as follows:

The subjects to be evaluated await singly, or in groups of six or more, outside the consultation room. The form and a pencil are handed to each by a civilian secretary, and the examinee fills out the Index, in

* The opinions stated herein are those of the author, not necessarily those of the U. S. Navy Dept.

from five to fifteen minutes. (Those taking notably longer are reported by the attendant, and are usually found to be recalcitrant, blocked, illiterate, or of limited intelligence.) Until recently, it was my custom to rate the form with the use of the three stencils for Form C or the single stencil for Form N, which required from a minute to a minute and one half. It was found that a secretary can be easily trained to rate these forms, saving additional time for the psychiatrist. A rapid survey of the first page of the form gives an impression of the educational attainments, a very rough idea of intelligence and a summary of significant job preferences, self-esteem, and the number and relative importance of psychoneurotic determinants. Essential facts of the examinee's history and complaints, if any, are recorded on the front page of the form, including such information as draft status, military service (combat or non-combat), etc., before making an inspection of the significant answers on pages one, two, and three. Additional information on any point is then obtained and recorded on the front page. Detailed histories are taken when indicated, often supplemented by letters from family physicians and civilian psychiatrists.

The relative value of Form C and Form N is debatable. The former gives much additional information: *e.g.*, in the statement of job preferences, the examinee is likely to bring out manifestations of homosexuality not otherwise readily discoverable. Likewise, on page two of Form C, the patient's own estimate of himself is valuable in determining the degree of uncertainty and insecurity under which he labors, the extent of his feeling of inferiority or over-compensation. While pages one and two contribute comparatively little, quantitatively, they are reliable, qualitatively. The aggregate information seems to be far greater than that obtained from Form N alone. However, in sixteen year old boys applying as apprentices, use of the single sheet Form N saves considerable additional time and gives quite adequate information. Infrequently do these youths have sufficient conflicts to be manifest, perhaps because they have not yet been confronted by life's problems. (A lesser number of significant answers in a sixteen year old youth probably should be considered suggestive of clinical psychiatric states than in older age groups.)

Statistical conclusions can hardly be drawn from this use of the Index, inasmuch as only those suspected of having psychiatric conditions were examined, except among applicants as police guards and sixteen year old apprentices. However, it can be roughly stated that the adult showing a score of fifteen or more on page three of Form C or

on the single page of Form N had to be viewed with suspicion. Those showing a score of twenty-five, or more, invariably fell into the category of severe psychoneurotics and could, therefore, be earmarked for more careful questioning. Those showing scores of less than fifteen could almost as readily be accepted for employment, unless there were significant answers to the "stop questions" or unless attention was called to the few psychiatric conditions not indicated on this Form. As pointed out by Weider, Mittelmann, Wechsler, and Wolff, many persons who are ill may not be detected by the use of the Index alone, but those detected *are* ill.

Psychopathy is not likely to be detected by this Form alone, unless accompanied by a psychoneurosis, or a history of reform school, arrests, alcoholism, or frequent changes of employment, which may serve as a guide to further questioning

Rarely do patients fail to answer truthfully the question, "Were you ever a patient at a mental hospital?" On the contrary, they have usually been so anxious not to be caught in written untruths, that they are likely to answer "Yes" to that question, whereas, actually, the confinement was to a general hospital for injury or non-psychiatric observation. Likewise, the question, "Have you ever had a fit or a convulsion?", is also answered with meticulous truthfulness, thereby leading to a diagnosis of epilepsy or hysterical loss of consciousness, equally undesirable in Yard employees or Navy personnel. Often, in his anxiety to answer the question correctly, the applicant is led to include such occurrences as "fits of temper" or simple syncope, also important in his evaluation.

The Index is equally applicable to females as to males, it being only necessary to reverse the significance of job preferences on page one. The average woman will give significant answers in from nine to fifteen job categories, as picked up by the stencil, whereas the average normal for males is less than nine, thus helping to validate the significance of the list of questions.

Perhaps the best way of illustrating the application of the information furnished by the Form is to quote individual cases.

Case 1—A handsome, intelligent twenty-four year old private, first class, United States Marine Corps, in service for two years, requested psychiatric examination because he was "fed up. disgusted, and worried about his widowed mother." He wanted release. He stressed the need to get back to look after his mother more than his other complaints,

Service history included a total of 210 days of hospitalization, 150 days of which had been for recurrent fungus infection of his feet.

He stated that he had been greatly attached to his mother, now widowed, and that she was suffering from a moderate degree of arthritis, which made him feel that he should get out of the Marine Corps and devote his time to her. He exhibited a typical letter from her, in which she dwelt upon her complaints. her loneliness and her need for him. He felt that her needs and desires were far more important than his duty to his country. He had worked for a radio broadcasting company (against his mother's wishes), and had served in the Merchant Marine for a year, prior to enlistment in the Marine Corps, and prior to our entry into the war. He had shipped over for another voyage after his first ship had been torpedoed. Obviously, he was not motivated by timidity in his present request.

His Cornell Selectee Index revealed eight significant answers in job preferences on page one, including a liking for dancing and singing and a dislike for aviation, carpentry, or the work of detective, forest ranger, iron worker or policeman, with the use of the stencil. It was also noted that he expressed preference for being actor, artist and music teacher. These additional feminine "likes" seemed more significant than his dislike for various occupations that are considered masculine.

Page two revealed twenty-one significant categories in which he considered himself poor, including keeping engagements, making decisions, repairing things, adjusting to strange places, making friends, getting a job, holding a job, impressing the opposite sex, controlling his temper, and mixing with strangers. He considered himself good in only twelve of the fifty-two categories.

Page three of the Index revealed forty-seven significant answers, including: fatigue, anorexia, nervousness, cold hands and feet, headaches, palpitation, nightmares, insomnia, frequency of urination, irritability, shyness, depression, unusual fears, sweating without exercise; often being misunderstood, a feeling of being watched while at work, of being watched or talked about in the street, of uneasiness when urinating in a public toilet, inability to make friends easily; nausea, faintness, difficulty in breathing, vertigo, fainting attacks, severe itching, indigestion, being upset by the sight of blood; and a history of a nervous breakdown, of drinking more than two quarts of whiskey a week, and of attacks of asthma. Of these significant answers, two were in the category of "stop questions," *i.e.*, nervous breakdown and excessive consumption of alcohol.

The extremely high total of forty-seven psychoneurotic determinants, or the significant answers to two "stop questions," would serve to call close attention to this man for further questioning His low self-esteem and his feminine job preferences were additionally significant. The impression of a severe anxiety neurosis was inescapable and that of homosexuality very suggestive. It was felt proper, in this instance, to ask him when he had first indulged in homosexual practices. He denied any such implication at first, but then stated that he had first noted his attraction toward males while working in a radio broadcasting station at the age of sixteen. He had had many promiscuous homosexual affairs, but claimed to have had a greater number of heterosexual relationships. He had been laid up for a month, at the age of eighteen, with marked anxiety symptoms. When asked if he did not realize the sequence of events: that his undue devotion to his mother was an evidence of immaturity; that the conflict set up by his homosexuality had induced or aggravated an anxiety neurosis (especially while in the Marine Corps); that this, in turn, had induced excessive sweating of hands and feet, which prevented complete cure of the fungus infection of his feet, he manifested ready understanding. It is doubtful that this information would have been obtained in so short a time, or, indeed, obtained at all, by a casual psychiatric interview.

Follow-up examinations, initiated by execution of an additional form, help to prove the value of the written questionnaire, as exemplified by the following:

Case 2—A twenty-three year old negro, discharged from the Army after twenty-one months of service, with no combat or foreign duty, because of "nervousness, headaches and dizzy spells," applied for a job as Helper Machinist. His answers on Form C all fell within a normal range, except for an affirmative answer to the question, "Did you ever have a nervous breakdown?" He explained this by stating that that referred to his slight nervousness in the Army. His only other significant answers were that he dreamed a great deal, and was often misunderstood. He had held a job as chauffeur for three years prior to service. He was recommended for employment with a diagnosis of "no definite psychiatric condition." Three weeks later, he requested limited duty with no work on ships, and gave quite a different story of having been extremely sensitive to noise and of amnesia for a period of six days while in the service. His answers on pages one and two of the Index were quite similar to his previous ones, but on page three his score of three had increased to thirty-two psychoneurotic de-

Case 6—A nineteen year old Marine had been absent over leave for one and one-half years, when he was picked up by detectives. He stated he had gone over leave because he had been in the brig and no one had paid any attention to his headaches. On further questioning, it developed that he had been in the brig because of having been A. W. O. L. The Index revealed thirty-one psychoneurotic determinants. He had had more than five jobs in two years. On following up this question, it developed that he had never been able to keep a job, that he had trouble throughout his school life, and that he had been arrested for rape. His family history revealed the common unhappy factors, including nervousness and illness in the father and mother, a crippled brother and three other brothers and two sisters, none of whom were adequate. He could not comprehend that his behavior in the Marine Corps was not normal or desirable. Survey from the service was recommended, as a constitutional psychopathic inferior with psychoneurosis, mixed type.

Case 7—A twenty-six year old male, 4F, following medical discharge from the Navy, was seen for psychiatric examination on application for employment. He had been in the Navy four months. When asked why he had been discharged, he said: "Sir, *I* had no complaints; they were complaining of me." He had attained the sixth grade in school at the age of sixteen, and had been unable to keep jobs He had laboriously filled out his Index (taking well over one hour). His answers pointed toward a mild anxiety neurosis. He was recommended for trial of duty as a laborer, with a rough psychometric evaluation of high-grade moron. A week later, he was arrested by the Yard police for bringing a can containing an unknown material into the Yard. When he was searched, a large slingshot was found. Considerable furor was created, until it was found that the can contained paint, that he had picked it up along the waterfront, intending to mail it to his mother, because it was associated with the area from which convoys sailed for Europe. He intended to use the slingshot for self-protection, because he felt that he was being followed at times by the men whom he had intended to shadow, while acting as amateur F. B. I. agent. It was still hoped that some job could be found for him, as a veteran. A week later he was reported to be springing in and out of windows, instead of using the doors, and it was felt that the Yard could no longer tolerate his schizoid, childlike and moronic behavior.

Use of the Index does not preclude malingering. However, it can often be readily detected by the inconsistency of the replies, which may

indicate a dislike for all jobs listed or poorness in all categories, with no other evidence of a psychiatric condition; or by all negative or all positive answers on page three of the Index. If the malingering is not suspected from inspection of the original record, a follow-up examination is likely to call it to the examiner's attention. For example:

Case 8—A twenty-five year old employee of the Yard was first examined on 29 September, 1944, after he had been employed for eight months. He stated that he was nervous, had low blood pressure, and wanted day work only. First inspection of the Index suggested a mild anxiety neurosis, while questioning brought out nothing of additional significance. Physical examination, previously made, was entirely normal. He was found fit for full duty. Four weeks later, he reappeared with the same request. His execution of the Index on this second occasion bore no resemblance to that of four weeks before, in any category. His job likes had decreased from nineteen to five; the categories in which he considered himself poor had increased from zero to twenty-one; and the psychoneurotic determinants had increased from seventeen to thirty-one. The only thing he was consistent about in his replies was frequency of bowel movements. It was apparent that there was, or had been, positive or negative malingering. He was considered unfit for further duty in the Yard and given a medical release. Five weeks later, he reapplied for the same job as a sheet metal worker. This time his replies to the written questions bore no resemblance to the two previous forms. His psychoneurotic determinants had decreased to one, apparently in response to his desire to have his job back, which might have been influenced by the fact that he was twenty-five years old and subject to draft. It seemed safer not to reemploy him, with a diagnosis of anxiety neurosis plus malingering.

Case 9—A twenty-two year old male applied for a job as Clerk, following discharge from the Army, after three months service, because of "deafness" and "nervousness." A diagnosis of psychoneurosis, anxiety neurosis, was made from information recorded on the Cornell Selectee Index and from questioning. There was some suspicion of a larger than usual homosexual component based on his appearance, mannerisms, and his preference for the occupations of actor, dancer, photographer, music teacher, nurse, school teacher and operator (only two of these preferences were picked up by use of the stencil). His self-esteem seemed disproportionately high in comparison with his complaints, humility, and mannerisms, with only five significant "poors" on page two. There were but five significant answers on page three of the In-

dex—nervousness, worry about health, spells of exhaustion and fatigue, enuresis between the ages of eight and fourteen, shyness. He seemed fit, as screened.

After working for two months, he requested psychiatric reexamination because he was excessively worried about being "all run down" after having had sexual intercourse two days previously. He added that he was worried about everything, but hastily added, "but not about intercourse—that is the cause of my physical exhaustion but not of my nervousness." He added: "I'm a good boy—my mother wants me to be a good boy—I'm a Catholic boy—I try to be good—I'm all right—I need to have a wife—I get upset about everything—I'm worried." He showed marked scattering, tension, confusion, blocking, and depression. He was too upset to safely question him further about suspected delusions, hallucinations, etc. He had executed a second form C with answers quite different from those of 27 October, 1944, with six "likes" on page one as compared with nineteen previously, the same number of "poors" on page two as previously, and with an increase from zero to fifteen "questionables." His psychoneurotic determinants on page three had increased from five to eighteen, including one significant answer to a "stop question." He modified many of his answers by writing in, "because I needed to bathe," after his affirmative answer to the question, "Are you at times bothered by severe itching?" and a large "'NEVER" after the negative answer to the question, "Were you ever a patient at a mental hospital?" A tentative diagnosis of incipient schizophrenia was made, and he was referred to the Veterans Bureau for treatment. He was denied admission, because it was considered not to be a service-connected condition. He returned on 8 January 1945, four days later, stating that he felt all right and wanted to go back to work. The form executed on that date showed similar answers to that of four days previously, but his psychoneurotic determinants had decreased from eighteen to twelve. He was advised to take more time off and to see a psychiatrist. He returned 5 February 1945, stating that he was improved and wanted to return to work. He showed no outward evidence of neurosis or psychosis. His fourth form C was similar to the previous two, except that there had been a further decrease in the number of psychoneurotic determinants, of which, this time, there were but seven, none of them essentially significant (*e.g.*, he no longer had the feeling that people were watching or talking about him in the street, etc.). He had been cooperative in all interviews, and his answers could be accepted as truthful. He was allowed to return to

work, but with the definite conviction that the noise and confusion of the Yard, plus his other problems, would probably precipitate further symptoms. A week later, he became frankly psychotic: belligerent, assaultive, and with marked disparity between mood and stream of talk.

Rather definite confirmation of the value of this combined method of psychiatric appraisal was obtained when a group of 40 men, screened from a total of 50 men sent for consultation by a medical officer, were referred for observation and disposition to a psychiatric ward of a naval hospital. The diagnoses, including psychoneuroses, psychopathies, and psychotic trends, were confirmed in all forty of these men.

SUMMARY

Use of the Cornell Selectee Index for more than nine thousand persons, civilians and naval personnel of both sexes, is described. The manner in which this method saves time, and allows one psychiatrist effectively to perform the duties ordinarily requiring at least two, is illustrated by brief case histories. The advantages and disadvantages of the written questionnaire are mentioned.

CONCLUSIONS

1. Use of the Cornell Selectee Index is an aid to the psychiatrist in forming a tentative, but usually accurate, opinion in a short space of time.

2. Obviously, it is not a substitute for a complete psychiatric examination, in itself.

DISCUSSION OF THE PAPERS

Lt. Comdr. William A. Hunt *(H(S)USNR)**: The question of malingering on these screen tests has arisen. Malingering does not present a serious problem. It is of two types—asymptomatic and symptomatic. The asymptomatic type occurs in normal, well-adjusted individuals under favorable environmental circumstances It is rare and easily handled when it occurs The symptomatic type, which occurs more frequently, is found in maladjusted individuals and is a function of some basic personality disorder. Attention should be directed, not to the malingering, but to the fundamental psychopathology. The significant thing is not that the subject malingers on a screen test, but that his malingering calls him to the psychiatrist's attention, and thus results in a recognition of the personality disorder responsible. Malingering is easily detected on screen tests. The subject invariably attains a much higher score than does the non-malingering patient. We have demonstrated this experimentally.

In using such screen tests, the Navy does not consider them as an absolute criterion for rejection, but uses them as a preliminary, coarse screen to pick out a

* The opinions or assertions contained herein are the private ones of the writer and are not to be construed as official or reflecting the views of the Navy Department or the Naval Service at large.

limited group of high scorers for an individual psychiatric interview. Such a group contains about 25% of the men originally tested, most of them being false positives, but it also contains 80% of the ultimate rejects. It is the duty of the psychiatrist to separate the unfit from the false positives. His judgment is the final criterion.

The screen test will detect 80% of those men who ultimately will fail during their recruit training. This compares favorably with the 85% that can be picked up by a brief psychiatric examination alone. However, where we have found the psychiatric examination to have a false positive rate of only 2% of the total group examined, the false positive rate of the test is approximately 25%. This difference seems largely due to the fact that the test questions receive a flat "yes" or "no" answer which cannot be qualified or expanded, whereas the psychiatrist is able to follow up answers which seem to him to need further elaboration.

It is understandable that such inventories should work better in a military than in a civilian situation. Because of the penalties imposed on lying in the military services, the subject is motivated to tell the truth. Moreover, the environmental demands in the services are so severe, and the allowable behavior patterns so restricted, that it is relatively easy to define the factors underlying maladjustment. Thus, enuresis, epilepsy, psychoneurosis, and psychosis all preclude adjustment in the Naval Service, but need not preclude adjustment in a civilian environment. Finally, in civilian testing, we usually aim toward an analysis of a complicated personality structure; in military screening, we aim only toward the prediction of adjustment to the service. While the problems of military screening are exceedingly vital for a society at war, they are basically much less complicated than those of personality testing in a society at peace.

Lt.-Col. Morton A. Seidenfeld (*Chief Clinical Psychologist, War Department, Washington, D. C.*): I have become acutely aware of certain difficulties that arise in the use of the Cornell Indices. For example, it seems quite apparent that whoever makes use of these or similar instruments must be prepared to make adjustments in the caliber of the screen to meet the characteristics of populations as they vary in different geographic areas. Furthermore, it seems altogether likely that similar adjustments need to be made, from time to time, as a result of the withdrawal of certain segments of society into military service, thus leaving a residual group whose mental characteristics are somewhat altered. It would seem that optimal selection will occur only when the characteristics of the screen are maintained on a dynamic basis, with alterations being made as frequently as the needs of the military service require them.

There is no question in my mind that Commander Harris made use of the Cornell Selectee Index merely as one of a number of psychometric devices which supply supportive evidence for both the psychiatric diagnosis and for further exploration of the deep-seated problems of the patient. However, considerable caution must be observed lest the implication be given that this and similar instruments have, within themselves, adequate information for diagnostic purposes. It is unfortunate, perhaps, that certain patterns of response on indices and questionnaires tend to become associated with specific mental deviations to which definitive names are applied. However, their values appear to be great as a starting-point for psychiatric exploration, in spite of the shortcomings indicated.

Dr. Harold J. Harris (*New York, N. Y.*): The Cornell Selectee Index, as I stated, is an adjunct to a psychiatric examination, not a substitute for it. The form, alone, is not relied upon in arriving at a diagnosis, but often brings out information not likely to be obtained otherwise, in addition to saving a great deal of time in the oral interview.

Lt.-Col. N. W. Morton (*Directorate of Personnel Selection, Army; Ottawa, Canada*): I suggest that the term, "personality test," as applied to various tools here described, is something of a misnomer. When we use the word, "test," we often think in terms of a scale measuring some definable trait common to the population tested, and existing in varying degree among different persons. Such a con-

cept, corresponding perhaps to Allport's definition of a nomothetic trait, may be unrealistic in relation to the appraisal of the largely non-cognitive qualities of personality; hence, the failure of measurement devices such as the Bernreuter Inventory. On the other hand, instruments now in use in the armed services to assist the screening and rapid examination of individuals of doubtful personality qualities are characterized by the fact that they do not aim to measure polarized traits, but rather serve to indicate roughly the extent to which the individual's health problems would be of significance in relation to service. This seems to me to be a very different thing, more readily approachable, rooted in concrete experience rather than in abstract definition of personality variables. The questionnaire or inventory works merely because it refers to many of the same items about which the clinician would himself enquire in an interview

On the subject of malingering in relation to military service, I may add that whether malingering is in itself symptomatic of maladjustment may in part depend upon the accepted social norm It assumes, I think, general good motivation and willingness to serve. If, on the other hand, the population includes sizable elements, distinguished by cultural background, which do not support this assumption, malingering may not be treated quite so simply in terms of individual maladjustment It may, indeed, be indicative even of strong conformity to a group pattern.

NON-PROJECTIVE PERSONALITY TESTS

PART III

ABILITY PATTERNS AND PERSONALITY

THE EXPRESSION OF PERSONALITY AND MALADJUSTMENT IN INTELLIGENCE TEST RESULTS*

By Roy Schafer

The Menninger Clinic, Topeka, Kansas

A. INTRODUCTION

Intelligence testing has long remained a technique for determining mental ages or I. Q.s, but has remained unconcerned with the impairments of specific intelligence-functions called into play by the different test-items. For example, it has remained unconcerned with whether the vocabulary score is high and the learning efficiency score low, or *vice versa*, as long as the total score remains the same. This gross approach leaves intelligence testing a technique in no way helpful in assessing personality and maladjustment.

In recent years, interest has begun to swing to scatter analysis, that is to say, to the analysis of the distribution of passes and failures within the test, or the unevenness of achievement in the test as a whole. For the most part, these studies have been concerned with finding a good gross quantitative measure of the amount of scatter or unevenness within the test. But even these procedures do not advance intelligence testing as a technique for exploring personality and maladjustment. At best, they can demonstrate a statistical trend, which may be of theoretical significance or may lend itself to mechanical-diagnostic attitudes, but which does not relate, in any palpable way, to the characteristics and individual variability of any specific case.

In order that intelligence testing may become a technique for detecting manifestations of personality-organization and maladjustment, a number of assumptions have to be made and thought through, about intelligence-functions and intelligence tests.

First of all, it is necessary to recognize that, in thinking about any individual's "intelligence" for diagnostic purposes, one must think in terms of a variety of intelligence-functions such as judgment, anticipation, concentration, etc.

---

* The procedures and results referred to in this paper are drawn from a complete report of a study of 7 psychological tests,[1] directed by David Rapaport, Ph D , and sponsored by the Josiah Macy, Jr , Foundation and the Menninger Foundation.

Secondly, it is necessary to recognize that different intelligence-functions underlie achievement on the different item-groups in an intelligence test, and that, consequently, it is necessary to establish what each of these functions may be.

Thirdly, it is necessary to recognize that the development and efficiency of each of these intelligence-functions are integral parts of the individual's personality development, and are regulated in their development by the vicissitudes of his needs and drives with all their emotional derivatives. From the time of Kraepelin[2] up to today, standard clinical descriptions of the different psychiatric disorders or personality types have always referred, explicitly or implicitly, to manifestations of symptomatology or modes of adjustment in the person's thought processes and intelligence-functioning. However, these attempts have been, in the main, descriptive, and have not related the development and efficiency of intelligence-functions to the dynamics of specific illness or personality. Specifically, the patterning of intelligence-functions does not relate directly to needs, wishes, and drives, but rather to preferred modes of control of these, or limitations of the expressions of these It is by these modes of control that we know the individual However, this general thesis must be made specific by recourse to data on the pathological conditions under which each of these functions becomes profoundly impaired, temporarily impaired, remains well-retained, or is even heightened in its scope and efficiency. An integration of these data with the psychiatrist's descriptions of the dynamics of the cases studied will then become the condition for inferring personality characteristics or maladjustment from intelligence test results.

At this point, intelligence testing may appear to become a projective technique, in the broad sense that it can be seen to elicit expressions of personality through the medium of the intelligence-functions. However, intelligence testing remains an essentially non-projective technique, in that it does not use the medium of individual organization, manipulation, and elaboration of *unstructured* test materials The subject must cope mainly with directly meaningful material and explicit requirements, and he may safely fall back upon verbal stereotype and memory, unlike any of the standard projective tests.

Finally, for using intelligence tests as non-projective tests of personality, it is necessary to recognize that there exist, in the general "normal" population, trends to have specific relationships between achievements on different kinds of subtests These relationships should be considered norms, deviations from which are significant for

the individual case. If Z-scores or derivatives of Z-scores are used to denote achievement on each subtest, then discrepancies between the subtest scores in any individual case become crucial diagnostic data.

Consequently, it is necessary to devise quantitative methods of analysis of intelligence test results, in order to explore and refine, with their help, qualitative methods of analysis. These quantitative methods not only utilize the test results to their full diagnostic advantage, but render them "objective" and easily communicable. The intelligence test, of choice, is one which lends itself more readily to such analyses. We chose the Wechsler Bellevue Scale.[3] Before describing our methods of analysis, it will be necessary to describe briefly the test itself.

B. THE BELLEVUE SCALE

The Bellevue Scale comprises eleven subtests, each containing relatively homogeneous, but increasingly difficult, items.* The subject's achievement on each subtest obtains an independent score from 0 to 17; these scores are equated-scores, that is, they are derived from Z-scores and are therefore intercomparable. Thus, a score of 15 on one subtest and a score of 9 on another indicate a definite superiority of the development and efficiency of the function underlying achievement on the former. The scale includes six Verbal subtests, counting the Vocabulary subtest, and five non-verbal or Performance subtests. Recognition is thereby taken of the fundamental differences between the thought processes underlying achievement on items requiring verbal, and those requiring visual and/or motor, performance

C. METHODS OF ANALYSIS

Personality and maladjustment can be traced in the intelligence test results from 3 points of view, *i.e.*, by three main methods: 1. In terms of relative impairment or superiority of a function; 2. in terms of its temporary inefficiencies; 3. in terms of the subject's manner of verbalizing his responses.

(1) Impairment or Superiority of Functions

The impairment or superiority of the different intelligence-functions involved in the intelligence test is established by scatter analysis. Scatter analysis, as used here, is analysis of the quantitative differences between the weighted subtest scores of the Bellevue Scale. Several

* The only exception is the Digit Symbol subtest.

different types of comparisons are possible: a comparison of the Verbal score-level to the Performance score-level; a comparison of the achievement on any Verbal test to the general level of the remaining Verbal tests, and similarly for the Performance tests; a comparison of the achievement on subtests vulnerable to maladjustment with the relatively sturdy Vocabulary score; and intercomparisons of specific subtest scores, which clinical experience has indicated to be a fruitful source of diagnostic indications.

Scatter analysis is dependent upon establishing baselines from which to estimate impairments or superiorities; our experience shows that three main baselines may be profitably used:

(a) *Vocabulary Scatter.* Vocabulary Scatter is based on the well-known finding[4] that, of all intelligence test scores, the Vocabulary score offers the greatest resistance to impairment by maladjustment. In almost all clinical cases, it is the Vocabulary score from which the premorbid level of intelligence development—before pathology impaired intelligence-functioning—can best be inferred. The remaining scores show greater or lesser vulnerability to maladjustment, and, therefore, comparison of these scores to the Vocabulary score as a baseline will indicate the extent of impairment. A slight amount of variability of the subtest scores below the Vocabulary score is frequent even in the normal range, but in this range the Comprehension, Information, and Similarities scores, in general, stay on the same level as Vocabulary.

(b) *Mean Scatter.* It is also necessary to see how the more vulnerable scores compare to each other. Mean Scatter measures the difference between the score on any subtest and the general level of the scores on the remaining subtests. Thus, if all the scores excepting Vocabulary are pushed down markedly, Mean Scatter may demonstrate that the Comprehension score, for example, is especially impaired, as it is, frequently, in chronic schizophrenics. Analysis of Mean Scatter should be pursued separately for the six Verbal subtests and five Performance subtests.

(c) *Specific subtest comparisons.* Also crucial to scatter analysis are the specific comparisons of pairs of subtest scores other than the Vocabulary score, aimed at answering such questions as: "How does the subject's judgment compare to his fund of information?" "How does his capacity for attention compare to his capacity for concentration?" etc.

Thus, scatter analysis traces the relationship of the score on one subtest to the Vocabulary score, in order to estimate extent of impairment; to the general level of other scores, in order to estimate especial impairment; and to single other scores, the relationship to which is a crucial datum for the understanding of the personality and maladjustment of the subject.

(2) Temporary Inefficiency

Inasmuch as each of the Bellevue Scale subtests comprises homogeneous items of varied degrees of difficulty, item-analysis of the sequence of passes and failures within each subtest clarifies the smoothness and efficiency of the function or functions underlying achievement on that subtest. It is important to know whether the final subtest score derives from failures on the "easy" items and successes on the "difficult" ones (in which case we speak of temporary inefficiency), or whether the failures first set in on difficult items and mark the point where the subject's development is no longer adequate to cope with the new items. The difficulty of items can be statistically established by the incidence of failure on each in the general population

Inefficiencies may be due either to intense anxiety or to a psychotic process When anxiety is their source, the responses are characterized by uncertainty, false choice between the correct and an incorrect alternative, and quick or delayed correction of wrong answers Furthermore, these failures on the relatively easy items will tend to be few and to occur on those items which, although "easy," are frequently missed by other cases showing inefficiency For example, in the Information subtest, the capital of Italy may be Naples, the average height of American women may be 5 feet 2 inches, there may be 4 pints in a quart and 48 weeks in a year, etc. Intense anxiety can prevent knowledge or ideas, once acquired, from emerging into consciousness and may lead to false and inaccurate responses. However, anxiety, alone, does not account for answers so incorrect as to be absurd. Both of these, especially when accompanied by a degree of bland confidence, are indications of a psychotic process. Furthermore, if the subject knows what ethnology and the Apocrypha are, and yet does not know where Brazil is, this, too, is a psychotic indication: the discrepancy between retained and lost information here is too great to be accounted for merely on the basis of anxiety-determined, temporary inefficiency.

(3) Analysis of Verbalization

Thus far, we see that departure from gross statements about I. Q. is accomplished by scatter analysis, as described above; that scatter analysis is amplified by item-analysis, as described above; and now amplification of the test results, in general, is accomplished by qualitative analysis of the subject's verbalization of his responses. In verbalization, we can follow the subject's intelligence *at work* and, thus, we often see in it, in clear form, the expressions of the subject's maladjustment or of the main aspects of his personality make-up. This analysis is concerned with *how* the subject failed or passed the test-items; that is, were doubt and indecision characteristic, were more or less bizarre ideas expressed, was impulsiveness prevalent, etc.?

It must be remembered that, although the correct responses to the test-items are fixed by common agreement, the routes to these, as well as to the incorrect responses, are not fixed. Thus, the subject's manner of reasoning out Comprehension or Similarities or Arithmetic items, his speed and confidence in delivery of his responses, his anxiety or blandness about incorrect responses, are all revealing of him For example, when an otherwise intelligent subject states blandly that the capital of Italy is Constantinople, that a dog and a lion are alike because both have cells, etc., the diagnosis of schizophrenia is strongly indicated. When a subject lists five alternative courses of action or explanations on some of the Comprehension items, three similarities on some of the Similarities items, and gives extensive and quibbling definitions on the Vocabulary items, etc., obsessive character make-up or obsessive pathology must seriously be considered. This occurs in paranoid cases, too, but in these the circumstantiality or pedantry will become peculiar in places. If wild guessing occurs on every item, no matter how difficult and obviously beyond the subject's level of ability it may be, psychopathic trends are likely to be present.

Furthermore, on a number of the Bellevue Scale items, common agreement or consensus of opinion are not knowable to a subject and here verbalization may become especially revealing of him. This is true, for example, of the difficult Picture Arrangement items, where distorted anticipations may be verbalized in explaining the sequence of pictures offered. One paranoid subject saw a woman rejecting a man's attentions by signalling to another man, although it was the same man in both pictures and no indications of any signalling are present in the pictures.

In the course of experience with any test, an examiner becomes familiar with the usual forms of verbal expression used by subjects, and with the usual errors or failures. It is against this subjectively-retained experience as a baseline that he may detect deviant verbalizations, that is, verbalizations the form or content of which stems from a specific maladjustment type or personality-organization. To date, however, verbalization is the least explored and systematized of all the material elicited by intelligence tests, and, for that matter, by tests in general.

D. SAMPLE ANALYSES

Let us analyze, by the techniques outlined above, the Bellevue Scale records of a few cases. Their scores are presented in FIGURE 1.

CASE 1—OBSESSIONAL NEUROTIC

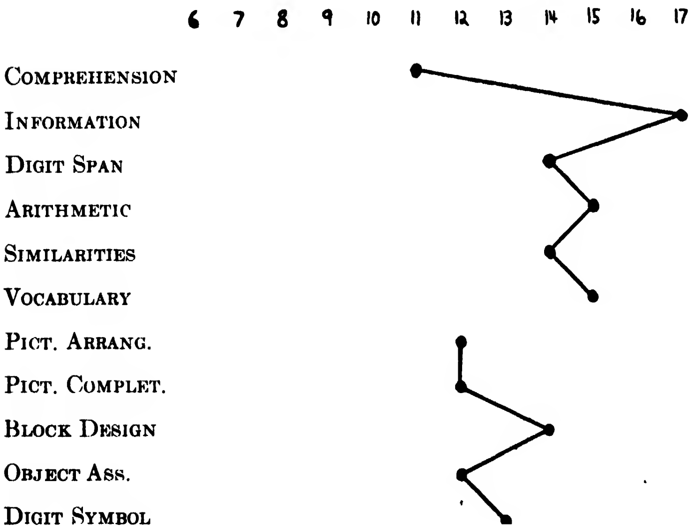


FIGURE 1

Case 1 is an obsessive-compulsive neurotic. If we look first at the scatter, its outstanding aspects are the impairment of the Comprehension score and the superiority of the Information score. In everyday experience, this pattern of verbal achievement is repeatedly encoun-

tered in the records of cases of obsessive character make-up or of those with obsessional symptoms; furthermore, our statistical analyses show this scatter pattern to significantly differentiate obsessive cases from other neurotics.[1] The rationale of this diagnostic pattern is the following: the superior Information score must represent an especial inclination to pick up facts, knowledge, general information; and hence refers to an intellectualizing mode of adjustment. This becomes confirmed in analysis of verbalization. Many responses of the following type of exhaustiveness occur: the subject defines *diamond* as "a carboniferous gem found mostly in South Africa, it is mined, it can be used as an adornment or on machine tools," and then asks confidently, "Do you want more?" Furthermore, the low score on Comprehension refers to an impairment of judgment, that is, of the ability quickly and definitively to make decisions harmonious with the objective as well as affective aspects of a situation. The form of appearance of this type of impairment of judgment is seen in the characteristically obsessive, doubt-ridden, and indecisive verbalization about any course of action or result of reasoning. In accord with the inclination to know all the facts and angles, too many alternatives come to mind at once, judgments are formed full of reservations, so much so that they do not really qualify as "judgments," and a final decision is hampered, if not altogether avoided. When asked, "What should you do if while sitting in the movies you were the first person to discover a fire?", the subject replied, "Either ring the fire-alarm, or tell the manager, or jump up and shout 'fire'." He was asked which of these he would do and he replied with distress, "I don't know which I would do: how honest can I be with myself?" All these test findings establish obsessive pathology for this case. That this is no obsessive, normal subject with poor judgment is seen in the verbalizations; that this is not a preschizophrenic or schizophrenic impairment of judgment is established by the absence of any irrealistic (peculiar or queer) verbalizations, by the results of item-analysis, and by the minimal scatter of scores other than Comprehension and Information. Further support for the obsessive neurosis diagnosis is seen in the slight lowering of the Performance subtest scores, reflecting the hampering of action by doubt, excessive caution, and meticulousness.

Case 2 is a conversion hysteric. If we look first at the scatter, its outstanding aspects are the impairment of Information and Digit Span; Comprehension, *i.e.*, judgment, is well retained. In everyday experience and in our statistical analyses, this pattern of achievement is

CASE 2—HYSTERIC

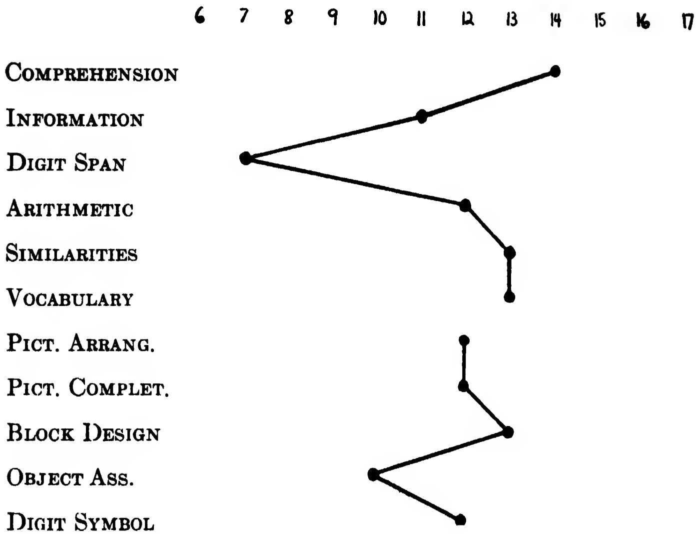


FIGURE 2

typical for hysterical neuroses.[1] The rationale of this diagnostic finding is the following: Digit Span, to our understanding, is a test of attention, that is, of the capacity for the free intake of stimulation. Digit Span does not correlate with tests of immediate memory, nor is its material meaningful, and, therefore, it is not considered a memory item by us. As a test of attention, Digit Span is especially vulnerable to impairment by anxiety. We have data to show that, even within the normal range, the Digit Span score progressively drops with increasing degrees of anxiety, as assessed by clinical observation.[1] Hence, our first inference about the case is that intense anxiety is present: there is a 6 unit difference between the Digit Span and Vocabulary scores. In regard to the low Information score, our understanding is this: at the core of hysterical neurosis is a pathologically excessive use of repression as a mechanism of defense. Effects of repression become widespread, when knowledge, information, extensive and speculative thinking all may become dangerous to the hysterical adjustment and, therefore, have to be avoided; otherwise, these thoughts might, in some more or less distant way, touch upon the repressed ideas and thereby mobilize further anxiety. As a result of the widespread repression,

the function underlying acquisition of information suffers, and the Information score becomes low. Item-analysis amplifies this point: it is not that all information suffers, but rather that there are sudden gaps in the information retained. For example, this subject knew who discovered the North Pole and who wrote Huckleberry Finn, but said that there are four pints in a quart, calculated that there are 48 weeks in a year, and could say of the heart only, "It beats." Analysis of verbalization shows that, in regard to the fire in the movie house, the subject would "yell 'fire' and run to the exit." The contrast of this impulsive response and the high Comprehension score indicates the characteristically hysteric-like affective liability. All these indications establish the diagnosis of hysteria.

Case 3—Neurotic Depressive

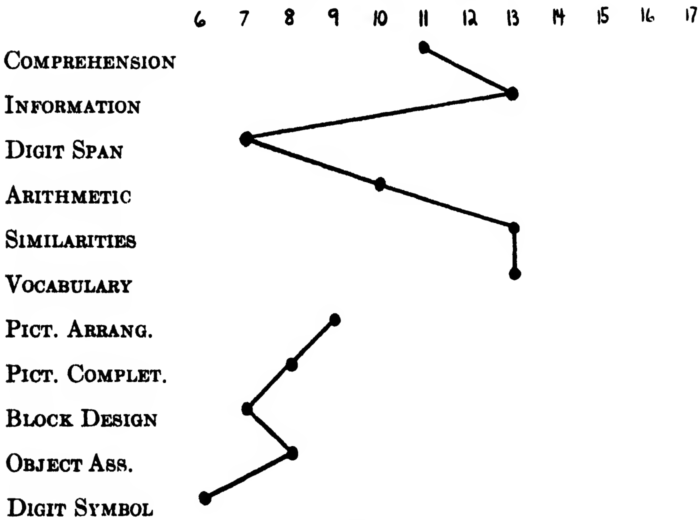


Figure 3

Case 3 is a neurotic depressive. Outstanding in the scatter is the great discrepancy between the Verbal and Performance subtest scores, and this discrepancy we have found to be a statistically significant, and therefore diagnostic, indication of depression.[1] The rationale of this finding is the following: depression becomes manifest in intellectual functioning, by a retardation of perceptual and associative proc- The relatively complex visual organization and visual-motor

coordination required by the Performance subtests put too great demands upon the slowed-down depressive. Furthermore, in contrast to the untimed Verbal subtests, the Performance subtests have time-limits on each item and even give extra credit for speed. Consequently, depressives not only do not obtain extra credit, but exceed the time limit on many items. For this case, item analysis confirms the retardation by showing that, on Picture Completion and Block Design, a number of items were failed only because they exceeded the time limit. To return to the scatter, the impairment of Digit Span or attention is also striking and reflects the presence of intense anxiety accompanying the depression. The mild impairment of Arithmetic is referable to an inability to meet the time-limits and gain time-bonuses on the items of this subtest. In verbalization, much self-deprecation, as well as indirect criticism of the test and examiner, are evident. The diagnosis of depression is clear.

CASE 4—SCHIZOID NORMAL

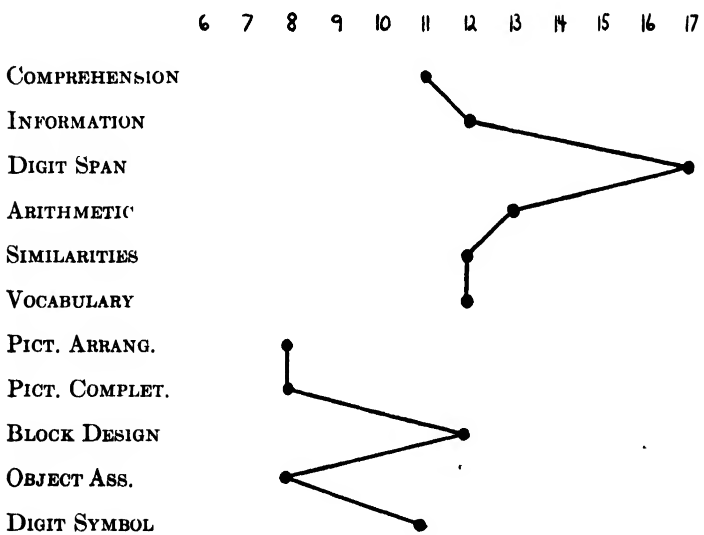


FIGURE 4

Case 4 is a normal subject who was judged on the basis of psychiatric interview to be definitely schizoid, that is, to show strong withdrawal tendencies and little interest in, or fellowship with, people of

his environment. Striking in the scatter is the great superiority of the Digit Span score; the overly-sharp attention indicated thereby we have found in our data to show a significant correlation with schizoid trends in the normal range.[1] The Performance subtests show more than the average amount of scatter, with Block Design being the highest score. This total scatter pattern is often encountered in the records of preschizophrenics and acute schizophrenics, but differential diagnosis here is made by recourse to item-analysis and analysis of verbalization. The absence of irrealistic thinking or performance, and the persistently good achievement on the easy items of each subtest speak against any acute pathology. Instead, we see a schizoid normal subject with excellent attention (Digit Span), some weakness of ability to make correct anticipations (Picture Arrangement and Object Assembly), and some impairment of concentration (Picture Completion).

These few analyses have been presented to illustrate the methods of analysis of, or ways of looking upon, intelligence test results which we found to be the most useful for the diagnostic application of these tests. To repeat: know the underlying functions, use scatter analysis, item-analysis, and analysis of verbalization. Knowing the functions underlying achievement on the different item-groups is the main safeguard against mechanical application of diagnostic scatter analysis. Mechanical application is a hazard, because intelligence test records are not always as diagnostically clear as in the few cases described above; there are cases whose scatter in no way reflects their diagnoses and even points away from the correct one. Special environmental-educational advantages or disadvantages, mood swings, or a generalized anxiety state accompanying the crucial diagnostic symptoms, etc., may all obscure the significance of the test findings. Furthermore, the maladjustment may find other avenues of expression than in the shaping of intelligence-functions. Due to specific conditions, an hysteric can have excellent information and poor judgment, an obsessional neurotic can have a fund of information inferior to his other achievements, a depressive can have great scatter with high scores on some of the Performance subtests, etc. If a battery of projective and non-projective tests is used, the atypical variations of scatter will not dismay the examiner, but will be used by him to draw out the specific flavor of the individual case. The other tests will establish the diagnosis.

E. THE CRITERIA FOR DIAGNOSTIC INTELLIGENCE TESTS

From this analysis, the following ideals may be derived for the development and application of an intelligence test as a non-projective test of personality and maladjustment: (1) The test must include homogeneous item-groups. (2) The intelligence-function or functions underlying achievement on each item-group must be known to the examiner, so that he can reconstruct a psychologically meaningful and differentiated picture of the subject's intelligence. (3) Within each item-group, there should be a carefully established sequence of increasing degree of difficulty of the items, so that item-analysis can be pursued. (4) The final score on each item-group or subtest must be translated into an equated score which will be directly comparable to the equated scores on the other subtests. (5) The range of the equated score-scale should be sufficiently wide to allow for representation of fine, as well as great, differences in achievement (6) One of the item-groups must be a Vocabulary test, because, thus far, the score on this has been found to be the most reliable indicator of the subject's potential, or premorbid, level of achievement. (7) The test must include verbal and non-verbal item-groups. (8) The specific item-groups, other than Vocabulary, must be selected to call into play intelligence-functions which are specifically vulnerable to, or specifically regulated by, different kinds of maladjustment and personality organization. This selection must be based on a scrutiny of clinical descriptions of the various psychiatric disorders for statements about differentially diagnostic impairments of specific intelligence-functions; and it must be based on a scrutiny of results, obtained with intelligence tests already extant, to detect the kind of items which are most effective in establishing differential diagnoses. (9) The items finally selected or invented must be tested on a large number of clinical cases of all the classical diagnoses, as well as on a large number of normal cases with different types of personality organization and maladjustment tendencies. In other words, the traditional ideal of using merely a large number of "normal subjects in general" must be abandoned, and the diagnostic intelligence test must be standardized and validated, not only on normal subjects in general, but on specific kinds of normal subjects and on clinical cases, too.

An intelligence test, meeting or approximating all these criteria, allows for the most fruitful scatter-, item-, and verbalization-analyses

and, therefore, allows for a reliable reconstruction of the subject's characteristic mode of intelligence-functioning.

One final ideal must be discussed here. Traditionally, the content of each item in an intelligence test has been selected only on its ability to differentiate levels of intelligence. Exactly *what* the examiner was asking and *what* the subject was reasoning about have not been held relevant considerations, except, perhaps, in the qualitative analyses pursued by the individual clinician. However, intelligence-functions do not operate in a vacuum, nor are they consistently on one level of efficiency. The questions the psychologist must ask himself are: Information about what? Good judgment when? Planning ability with respect to what? In other words, the content of the test-items must be selected to refer to different areas of ideation, and affect above and beyond the ability of these items to differentiate levels of intelligence. Such a selection would allow for considerably richer analyses of verbalization than are possible with the already existing tests.

To be specific: it is apparent that acquisition of information is largely a selective process. Some things we must all know, but others we know only if we care to, and remember only if we are able. Hence, a random selection of the content of test-items allows, for example, only for statements about the subject's fund of information *in general*, and allows little room for inferences about characteristic inclinations or modes of accumulating information. Therefore, tests of information should be designed in which information pertaining to a variety of different areas of ideation is measured. For example, the items might inquire about things sexual, aggressive, practical, abstract, esthetic, scientific, everyday, remote, etc. From the discrepancies between the achievements within each of these areas of information, personal interest and inclination could be inferred, and statements about the subject's information could gain a profitable degree of differentiation. These same considerations hold for vocabulary and for all subtests where the subject must cope with meaningful material. To take other examples, Picture Arrangement could present sets of pictures dealing with danger, with fear, with expression of hostility, with sexually-toned situations, with humorous situations, etc.; tests of learning efficiency (unfortunately lacking in the Bellevue Scale) could present a similar variety of material to be remembered.

Needless to say, a test or an experimental study, using such varieties of content, would offer invaluable material, not only for assess-

ment of personality-organization and maladjustment, but for the general psychology of thinking and of verbalization of thinking. It would offer data to show how the development of some intelligence-functions is speeded or retarded in general, or is speeded or retarded with respect to only specific conditions. It would also have data to offer on how the subject's characteristic mode and efficiency of using his intelligence-functions vary in different kinds of situations or in connection with different problems. The final picture of intelligence drawn from the material of such a test would be a living one, pertaining directly to personality-organization and maladjustment-type, and would not be a statistical skeleton of scores and score-comparisons. Non-projective testing of personality through the medium of intelligence tests will have come of age.

REFERENCES

1. **Rapaport, D., M. Gill, & R. Schafer**
   1945. Diagnostic Psychological Testing. Yearbook Chicago.
2. **Kraepelin, E.**
   1917. Clinical Psychiatry. William Wood & Co. New York.
3. **Wechsler, D.**
   1941. The Measurement of Adult Intelligence. Williams & Wilkins. Baltimore
4. **Babcock, H.**
   1930. An experiment in the measurement of mental deterioration. Arch Psychol. **117.**

DISCUSSION OF THE PAPER

Dr. Z. A. Piotrowski (*Columbia University, New York*): The interesting contribution of Dr Schafer was prompted by a very natural desire, in a clinical psychologist, to utilize the intelligence tests, and observations made during testing, as aids in psychological diagnosis The actual experimental data presented by the speaker demonstrate how much can be accomplished in this direction I would like to ask Dr. Schafer whether he tried to raise the discriminative value of the tests by dividing his subjects into groups of similar mental ages or similar total scores Many years ago, I found that the original form of the Stanford-Binet tests differentiated psychotics from non-psychotics much better when the subjects were divided into three groups; with mental ages below 13 years; with mental ages of 13 to 15 years; and with mental ages over 15 years* It was interesting to note that the significance of certain tests varied according to the mental age Thus, non-psychotics with mental ages below 13 years excelled psychotics of similar mental ages on the test of repeating digits; but in the highest group with mental ages above 15 years, the psychotics, significantly excelled the non-psychotics on the same test. Also, I wonder whether Dr. Schafer would agree to describe those test items which are relatively easily affected by pathological mental states (*e.g.*, emotional tension), as open-eyes tests; and those which are relatively unaffected, as closed-eyes tests. The latter can be solved even when the eyes are closed (*e.g.*, vocabulary), while the former require close visual attention to the immediate environment (*e.g.*, the digit symbols, or colored cubes).

* Objective signs of invalidity of Stanford-Binet tests. Psychiat. Quart. **11:** 623–636. 1937.

PERSONALITY AND DIAGNOSTIC EVALUATION BY MEANS OF NON-PROJECTIVE TECHNIQUES

By Edith Wladkowsky

Psychological Department,
Psychiatric Division, Bellevue Hospital, N. Y.

In the use of intelligence testing, psychologists have for a long time stressed the importance of considering discrepancies and quality of responses in test results While earlier studies concerned themselves primarily with certain psychometric signs as indicative of emotional maladjustment in general, more recent studies have stressed the importance of evaluating specific types of maladjustment from test data.

Because the nature of the Wechsler-Bellevue Scale is such as to make it especially adaptable for the study of patterning in clinical entities, a number of patterns have been worked out for several types of abnormalities.

Levi, in his work on patterning in adolescent psychopathic boys, found that the two most important signs for this group are: The performance score is higher than the verbal results; and the sum of the Object Assembly plus Picture Arrangement scores is greater than the sum of the Block Design and Picture Completion Test scores.

Rabin, working with schizophrenic patients, found a number of signs which differentiate this group from non-schizophrenics The verbal I. Q. is generally higher than the performance I Q However, this characteristic has been found to exist in most clinical entities studied, except in psychopaths and mental defectives, where the reverse is generally true. Other signs found for schizophrenics in the Wechsler-Bellevue Test are the following: The sum of the Information and Block Design scores is generally higher than the sum of the Picture Arrangement and Comprehension scores; the Block Design score is generally greater than the Object Assembly score; very low Similarities with high Vocabulary and Information scores are definitely considered pathognomic for schizophrenia; the interest variability is usually found to be marked, especially on the verbal parts of the examination.

Similar patternings have been worked out and described in Dr. Wechsler's book, *"The Measurement of Adult Intelligence,"* for neu-

rotics, organic brain disease cases, and for mental defectives. In all of these cases, patients who have been classified in specific clinical groupings tend to show differences in the various subtest scores or in their ability for certain types of performances.

While consideration of ability patterning is important, analysis of the qualitative aspects of responses to different subtests gives added valuable data which often help to make a differential diagnosis.

In evaluating personality make-up from test items, the adage, "one can judge a person more by his actions than by his words," certainly seems true. One can often differentiate impulsive individuals from those who are deliberate; those who give up easily when confronted by difficulties from those who show perseverance in the face of difficult situations. Does the individual tend to perceive the whole situation or does he react to details only? Do his reactions show purposeful movements or mere trial and error behavior? Does the individual react to details or is he oblivious of such elements in the test situation? The answers to these and many other such questions present interesting and valuable material in the judgment of the individual's personality make-up.

A few examples to show what qualitative results some of the individual Wechsler-Bellevue subtests may yield will suffice. For the Picture Completion Test, Dr. Wechsler says: "In a broad way, the test measures the ability of the individual to differentiate the essential from the unessential details." Thus, the schizophrenic's obliviousness to details often causes him to make a poor score on this test.

According to Dr. Wechsler, "the best feature of the Object Assembly . . . is its qualitative merits." Some individuals will immediately react to the whole, followed by a critical understanding of the relationship of individual parts. Other individuals will almost entirely react in a trial and error manner, in an attempt to complete this test. This latter method is often used by both neurotic and schizophrenic patients. However, while both groups are apt to display trial and error activity, the neurotic group, as Dr. Wechsler has shown, will ordinarily not display bizarreness in its production, while the schizophrenic patients will.

Considering some of the verbal tests, we again get qualitative differences which help in the evaluation of personality make-up. Thus, for the Similarities Test, both schizophrenics and those suffering from organic brain disease may make low scores. When we analyze their responses, however, we find certain differences. Patients with brain

disease may be unable to carry out the conceptual thinking involved in this test. The similarities which he will give may be superficial and not adequately generalized. Schizophrenics, on the other hand, may show negativism by their insistence in giving differences instead, or by contaminating adequate responses by adding something inappropriate.

One of the best verbal tests from which to study the quality of responses is, perhaps, the Vocabulary Test. This is especially true in cases of schizophrenia or schizoid types of individuals, whose expressions of ideas in the form of language contain elements of bizarreness and pecularities. Thus, one incipient schizophrenic individual defined "diamond" as "a shining rock; you can be cut by; a diamond is found in Brazil, South America; it is very expensive." The word, "nail," he defined as "a metal substance, sharp at the end and is used to hold things together." For the word, "plural," he said, "It means singular or plural; singular is one; don't know what plural means." The word, "cedar," according to this patient, "means to drink cider." When the word, "cedar," was again emphasized, he said, "Yes, I know, the word, cedar, it means to drink cider." These last two illustrations may show the negativism often displayed by this type of patient.

The results of these, as well as of all other subtests in the Wechsler-Bellevue Scale, must, of course, also be viewed in the light of the patients' cultural differences and differences in training and experience, as discussed in Dr. Wechsler's book. Thus, people with a considerable amount of formal education are likely to do better on digit symbols than those who have had little formal schooling. Bookkeepers and accountants will, of course, be expected to do better on arithmetical reasoning than those of equal intelligence who are not so engaged. Carpenters will usually do better on the Object Assembly Test than those who have not had such training at all. Cultural differences will be especially reflected in the scores on the Picture Arrangement Test However, after considering such individual differences as may be due to different training and cultural background, variations in scores and in the quality of responses often reflect personality make-up, or, in the clinical sense, a possible break-down of personality.

No attempt is made in this paper to make a statistical study of ability patterning and of qualitative results on the Wechsler-Bellevue Scale in the evaluation of personality make-up. Two illustrations will merely be given, from actual cases, to show what such results may indicate in schizophrenia or in schizoid personalities. While a number

of tests are ordinarily given to any one individual, only the results on the Wechsler-Bellevue Scale will be discussed in each case.

An 18 year old white boy, a recent high-school graduate, desirous of entering college, came to the writer for vocational advice. History, as given by the mother, disclosed that this lad is the only boy in the family of three children. The two younger siblings, girls, ages 14 and 12, have always been considered brighter than this boy and have continuously made better progress in school, and with less effort. Also, at home, this boy has always been considered more troublesome, inasmuch as he has shown an undue amount of negativism. No other behavior difficulties were noted.

As a youngster, he associated with other boys. From the age of 14 or 15, however, he apparently became more secluded, had no friends, and quarreled with his sisters more than previously. Although he found high-school studies difficult, he nevertheless took a college preparatory course. His drive to excel in school became stronger. He studied every spare minute he had, but, at the same time, apparently he was unable to concentrate, and excessive day-dreaming was noticed. During this period, he became more and more secluded, and has shown a terrific amount of sensitiveness in the presence of other people, especially those of the opposite sex.

The boy told the examiner that he must go to college in order to excel, and that his greatest desire was to become "great." He has had no goal for any specific accomplishment which would lead to life work. Although, as a youngster, he expressed eagerness to work for remuneration, upon his high-school completion he resented any kind of work which was offered to him. His father, a skilled laborer, found him simple jobs which he has invariably given up, because he could not get along with employers or co-workers, with the result that, during the summer months immediately following his high-school graduation, he worked in four or five different places. Employers complained that the boy was slow and clumsy; was unable or unwilling to learn simple mechanical tasks, and that he was quarrelsome and negativistic.

While the boy desired merely to attend college so that he could "become great," it was the parents who insisted on getting professional advice for him.

On the Wechsler-Bellevue Scale, the boy obtained a composite I. Q. of only 91, with a verbal I. Q., however, of 112, and a performance I. Q. of only 70. Comparing the various subtests, the following weighted scores were obtained:

Verbal Subtest Scores		Performance Subtest Scores	
Comprehension	8	Picture Arrangement	4
Information	12	Picture Completion	3
Digits	14	Block Design	8
Arithmetic	10	Object Assembly	5
Similarities	12	Digit Symbol	10
Vocabulary	11		

The patterning of these data suggests at least a schizoid type of individual, intermingled, perhaps, with neurotic traits. In his patterning, the only clear cut sign for neurosis is his performance on the Digits Test, on which he made a higher score on repeating digits backward than forward. Such results have been found to be characteristic of many neurotic people. The signs which suggest schizophrenia are the great discrepancy between the verbal and performance scores in favor of the former; the very low score on the Picture Completion Test; the higher score on the Block Design than on the Object Assembly Test; and the sum of the Information and the Block Design scores, which is greater than the sum of the Comprehension and the Picture Arrangement scores.

Analysis of the different test results, from the qualitative point of view, again points to characteristics found in schizophrenia. Quite a bit of intra-test variability, especially on the verbal material, was present. There was also a tendency to contaminate good responses. Thus, in the Similarities Test, to the question, "In what way are a wagon and a bicycle alike?" he said, "Both are used for transportation and they both have wheels." In the Comprehension Test, to the question, "What should you do if, while sitting in a theatre, you were the first person to discover a fire?" he replied, "Tell the manager; go to the fire department."

His reaction time was slow for those tasks in the performance parts on which he obtained low scores, and his reactions contained elements of bizarreness. Thus, for the Picture Completion Test, according to him, the food was missing for the pig, water was missing for the boat, the back was missing for the watch, and a few trees were missing for the sun.

Much trial and error behavior was noticed, especially on the Object Assembly Test. By means of such behavior, he completed the "man" and the "profile" within time limit. However, he obtained a zero score for the "hand," despite his constant activity on this problem, and the result to him was "an animal of some kind."

The following deviations in his definitions of words were found from what people usually give. A "diamond" is "jewelry, one part of jewelry which consists of a precious stone." A "nuisance" is "one who is considered unpleasant by other people, and by other people is the important part; a pest." "Fur" is a "kind of attire and it comes from animals. Rich fur is mink. It is hair and leather; top hair and bottom leather." A "cushion" is "something to sleep on; you rest your head on a cushion on retiring." For the word, "nail," he said, "You mean this?" pointing to his own finger nails, and then went on, "for building purposes, carpenters use it." And, finally, for the word, "seclude," he said, "to go off in your own corner; be in apathy, alone; a trapper lives in seclusion."

The examiner, feeling that this boy was too disturbed emotionally to profit from mere vocational advice, referred him to a psychiatrist The psychiatrist, at first, was under the impression that the boy was severely neurotic. However, after having seen him at weekly intervals, five or six times, the psychiatrist, too, felt that he was dealing with a schizophrenic individual.

The following is another illustration of a patient who was diagnosed as being basically schizophrenic, though the picture of schizophrenia was not clear cut.

A 24 year old white man was brought to the hospital with the complaint that, for a few weeks previously, he had been disturbed, crying and sobbing. About five days before his admission, he left his wife. He then complained of headaches, "jitteriness," and "needles all over."

The patient was discharged from the army in 1942. He claimed that a sergeant had made homosexual advances to him and that he had rejected him. The sergeant went to the commanding officer and accused him of being a homosexual. This led to quite an uncomfortable situation for the patient, inasmuch as no one would have anything to do with him. He denied being homosexual or ever having had such experiences, though he spoke in a rather effeminate manner.

The patient claimed that he felt very sensitive at not having received a discharge button. He kept this a secret from his wife, but upon finding it out, she took it quite badly, and accused him of being a coward.

The patient was well oriented in all spheres and no gross delusional material could be elicited. Some auditory hallucinations were present, though they appeared to be more in the sense of conscious voices and of vivid awareness of his own thoughts. He admitted some suicidal preoccupations, but denied ever having attempted it.

During his stay in the ward, the patient quieted down considerably. He seemed reluctant to return home, preferring to stay in the hospital. He then admitted having had homosexual experiences since the age of 18.

On the Wechsler-Bellevue Scale, the patient obtained a composite I. Q. of 109, with a performance I. Q. of 104, and a verbal I. Q. of 112. His weighted scores on the different subtests were as follows:

Verbal Subtest Scores		Performance Subtest Scores	
Comprehension	11	Picture Arrangement	10
Information	13	Picture Completion	7
Digits	11	Block Design	11
Arithmetic	10	Object Assembly	10
Similarities	12	Digit Symbol	13
Vocabulary	12		

The patterning of the subtests in this patient seems even, except for the relatively low score on the Picture Completion Test. His intellectual functioning, therefore, seems intact. However, disordered thinking is found in the most taxing verbal tests. Quite a bit of variability is found in his responses to the Comprehension and Similarities Tests, and there is a tendency to spoil adequate responses by giving them an element of bizarreness. Thus, for the question, "In what way are a banana and orange alike?" he said. "Both grow on large plants, but they are not both fruits. Banana is not a fruit." "Is it a vegetable?" "No, technically speaking, it is a perennial." To the question, "In what way are wood and alcohol alike?", he replied, "Contains—alcohol contains oxygen and wood contains carbon dioxide. Both are matters of space." To the question, "In what way are praise and punishment alike?", the patient replied, "Praise somebody in a sarcastic manner and can make it a punishment." To the question, "Why are shoes made out of leather?", the reply was, "Because of the substantiality of material; you cannot make it out of anything else."

Although both patients were diagnosed as schizophrenics, the patternings of their test results do not give the same picture in each case. The first of these two patients shows a disturbance in his intellectual functioning on the performance tests. The second patient shows no such gross disturbance. Aside from the fact, of course, that there are all kinds of schizophrenics, it is difficult to give more precise reasons why these two people differ from each other so markedly in their individual ability scores.

Considering the quality of responses of these two patients, both show deviations from what are considered normal or usual responses. The responses of both are variable and contain elements of contamination and bizarreness. So, while the first case suggested schizophrenia, both because of the general patterning and the unusual quality of responses, the second case was discovered by means of the unusual quality of responses only. These cases, therefore, also indicate the necessity of interpreting both the quantitative results on the various subtests, and the quality of individual responses within the subtests.

In concluding this paper, it might be well to question the use of the terms, "non-projective techniques," as against "projective techniques." In the final analysis, all psychological tests are projective techniques, inasmuch as they are all used for purposes of interpreting the individual's make-up. Such functions as learning, memory, and specific capacities are as truly aspects of personality as are drives, wishes, and emotions. From this point of view, any psychological test can be used as a projective technique, especially if the qualitative results are carefully considered and interpreted in conjunction with the quantitative findings. The illustrations given above indicate to what use intelligence tests, like the Wechsler-Bellevue Scale, can be put, if more factors than mere intellectual classification are considered.

However, in using such tests for differential diagnosis of clinical entities, we have been validating our results with the diagnoses made by psychiatrists. But we know that psychiatrists do not always agree on any given diagnosis. Psychologists are, therefore, in need of more objective criteria by which they can validate test results. This should be a valuable problem for future psychological research. With more objective validating criteria at hand, psychological tests will become even more valuable instruments for the study of personality deviations and disturbances than they are at the present time.

DIFFERENCE BETWEEN CASES GIVING VALID AND INVALID PERSONALITY INVENTORY RESPONSES

By Zygmunt A. Piotrowski

College of Physicians and Surgeons, Columbia University, N. Y.

PROJECTIVE VERSUS NON-PROJECTIVE

This is a conference on non-projective personality tests. When applied to experimental personality methods, the concept, "projective," can be used with at least two different meanings. The concept usually designates techniques in which the subject is called upon to give a definite interpretation of an indefinite situation (interpreting ink-blots, giving a series of free associations, making up stories, organizing play material, etc.). The subject, unwittingly, reveals various traits of his personality by giving some specific explanation of a material susceptible to many more different explanations.[4]

The word, "projection," may mean something more closely related methodologically to geometric projection;[2] *e.g.*, in stereometry, we can conveniently study a three-dimensional and not easily accessible object, when we project it on a small and accessible two-dimensional plane. The terms, among which the relations hold, are changed through the process of geometric projection, but the identity of relations between the terms is preserved. The system of projection has its own laws and, if these are adequate, it is possible to make discoveries and predictions about those parts of reality that have not yet been discovered or studied empirically. Thus, the astronomers successfully predicted the discovery of several planets, and the chemists, the discovery of several elements. It is such a system of projection, with its own laws and inner consistency, used as an intermediary step in drawing conclusions from experimental data, which distinguishes some "projective personality methods" (*e.g.*, the Rorschach) from the non-projective personality inventories. In the latter tests, the results of every individual examination can be interpreted only in terms of direct, descriptive, statistical data and, therefore, never can attain accuracy when applied to individuals. Statistics is a descriptive study of groups, and not of individuals. By non-projective techniques, then, we mean personality tests, the raw results of which are interpreted only by refer-

ence to statistical, descriptive data. In the projective methods, the raw experimental data are interpreted according to a system of principles of general validity, before conclusions pertaining to the examined individual are drawn from the experimental findings.

PROBLEM

In this communication, an attempt will be made to describe some of the differences between subjects who give valid inventory replies and those who give invalid inventory replies; *i.e.,* differences between persons who describe themselves rather adequately and persons who give an inadequate self-description. The inventory which has been used in this survey is the revised edition of the Kuder Preference Record, form BB.[1] This inventory was devised to obtain a record of a person's preferences with respect to a variety of activities. Scores are obtained in the following 9 areas of activity: mechanical, computational, scientific, persuasive, artistic, literary, musical, social service, and clerical. I have attempted to determine the validity of the Preference Record only in activities grouped under the general heading of social service. Frederic Kuder lists a number of occupations "whose duties appear to be consistent with the activities" tested by the social-service scale. The primary purpose of the Preference Record is to serve as an aid in vocational guidance. The assumption is made that, while the scores are not measures of ability, they "may indicate what sort of a thing a person enjoys doing." Occupations which a person scoring high on the social-service scale would presumably enjoy are: camp director, clergyman, personnel director, psychologist, physician, camp counselor, Y. M. C. A. secretary, sociologist, interviewer, dean, college personnel director, nurse, occupational therapist, sales manager, welfare worker, criminologist, playground director, policeman, scout leader.

SUBJECTS

A Kuder Preference Record was obtained from 18 university graduate students who were all preparing for the same type of social work, and who all belonged to the same religious group (Protestant). Eight were men and ten, women. All but two were under 30. None was born or reared in a large city. The intelligence of the entire group was superior: 7 obtained Wechsler Bellevue I. Q.s of over 130, 9 obtained I. Q.s between 120 and 130, and 2 obtained I. Q.s between 110 and 119.

The Kuder Record Results

All of the subjects revealed a very high degree of interest in social work, none scoring lower than in the 86th percentile, and only 2 out of 18 scoring lower than in the 90th percentile. Four obtained the highest possible score. All subjects but three obtained their highest scores on the social service scale. These three scored still higher on the scientific interest scale. All, without exception, scored lowest on the clerical work scale, none of them scoring higher than in the 8th percentile on this particular interest scale. In all cases, the next lowest interest was that in activities requiring the art of persuasion, or activities relevant to such occupations as salesman, advertising manager, etc. Apparently, the whole group disliked influencing a person to act against his desires. A large proportion obtained high scores on the scientific scale. Few obtained high scores on the artistic or literary or musical scales. In other words, the profiles on the Preference Record were quite similar for the group as a whole. The number of subjects was not large. However, they were matched in many traits and had very similar Kuder Record profiles. These similarities and matching compensate somewhat for the moderate number of cases.

Criterion of Differentiation

Subjects who showed by their behavior, as it was actually observed and as it was reported by them in the psychiatric interviews, that they enjoyed human contacts, and that they possessed positive interest in social service activities, were placed in the valid group There were 11 of them The other 7 were placed in the non-valid group. The presence of persons not really interested in working with people for the latters' benefit, in a group preparing themselves for lifelong social work, seemed due to external circumstances, such as pressure exerted by parents, or marriage to a person who was intensely interested in social service activities.

Psychiatric Observations

Dr. Irville H. MacKinnon, Assistant Professor of Psychiatry at Columbia University, examined the subjects. This writer is grateful to Dr. MacKinnon for permission to read the psychiatric notes, which were of great value in this study. These notes disclosed the following differences: Some members of the valid group placed emphasis on self-discipline, and, on the whole, were more decided concerning standards of conduct. The non-valid group was inclined to dominate without

any desire to teach or instruct others. They were much more frequently concerned with themselves than the valid group were. Members of the valid group did not hesitate to admit, and usually to show, their annoyance with others when the latter behaved in a manner disapproved by them, but they combined their disapproval with a tendency to teach others. They were more fearless in expressing criticism directly than was the non-valid group. In fact, this tendency was so strong in some members of the valid group that they tried to suppress it consciously. The non-valid group was more timid socially. Their criticism of others was expressed rather in the form of sarcasm. Several of the non-valid group admitted having been very close, psychologically, to their mothers. The majority of the valid group reported having been rather fearful in the past, but having struggled consciously to overcome their fear, in which attempt they succeeded to a large extent.

Rorschach Findings

All 18 subjects were given the individual Rorschach examination. A glance at the table will show that not many significant differences were found. The records were scored in the conventional manner,[5] except

GROUP	No. of Cases	No. Resp.	W	WS	D	DS	d	S	h%
Valid...	11	40.0	11.8	1.0	21.8	0.4	4.0	1.0	18.0
Non-Valid....	7	32.6	12.2	0.8	17.0	1.8	0.8	0.0	7.0

GROUP	M	MC	FM	m	c¹	Fc¹	F	Fc	c	FC	CF	C
Valid...........	4.8	0.6	4.0	0.0	0.0	1.6	18.2	3.8	0.8	2.6	2.2	1.4
Non-Valid......	1.6	0.0	2.8	0.8	0.8	1.0	13.0	5.6	0.6	2.8	3.0	0.6

GROUP	M:C	W:M	FC:CF:C	R(VIII-X):R
Valid...	5.4:6.2	11.8:5.4	2.6:2.2:1.4	0.37
Non-Valid.....	1.6:5.3	12.2:1.6	2.8:3.0:0.6	0.32

for the *chiaroscuro* responses and the non-human movement responses.[3]

The differences, indicated in the tabulated averages, suggest a better adjustment to reality in the valid than in the non-valid group. Evaluated according to available interpretive Rorschach principles, the difference in the M:C ratios implies that the valid group possesses a capacity for a wider range of psychological experiences and is less dependent, in its actions, on the environment than is the non-valid group. The desirable ratio of W:M in adults is 2:1. The valid group's W:M

ratio is practically 2:1, while that of the non-valid group is about 7½: 1. In other words, the desire to achieve something outstanding is far in excess of internal resources in the non-valid group. A very significant finding seems to be the difference in the percentage of human responses, *i.e.*, of responses with human content. The higher H% in the valid group indicates a higher genuine interest in people, as individuals with their own destinies and feelings, a greater realization of individual differences. The non-valid group does not seem to be as aware of individual differences as the valid group. There was no difference in the preference for either warm (red, yellow) or cold (blue, green) colors between the two groups. There was a qualitative, as well as a quantitative, difference in the M, or the human movement responses. M in which two people did something together, occurred in the valid group more frequently, while M in which only one person performed, occurred more frequently in the non-valid group. Examples of the first type of M are: "Social dance, jitterbug, some sort of a rhythm"; "two waiters pulling a container"; "couple of Moslems meeting together and playing patty-cake." Examples of the second type of M are: "Dancer in fancy costume walking toward you"; "a ragged girl in a hurricane, dejected, lets herself be blown"; "headless gorilla, legs awfully big, spread out, standing." The difference in the number of *chiaroscuro* responses was not great, but the non-valid group produced more of the *chiaroscuro* responses. The difference implied that the non-valid group was more inclined to intermittent depressive moods, and that its members were less certain of themselves than the members of the valid group.

CONCLUSION

Twenty-eight years have passed since, under the pressure of war needs, R. S. Woodworth introduced the personality inventory technique. This technique, one of the most popular in the psychological world, has been disappointing. It has given satisfactory results only under very favorable, special circumstances, and when used for limited purposes. The chief reasons for its lack of success as a dependable method of experimental personality analysis seem to be: (1) the relative inability of the average person to discover and to describe adequately his personality traits; and (2) his unwillingness to communicate the self-observations freely, especially in writing. This investigation of 18 intellectually superior graduate university students, all of whom scored very high on the social service scale of the Kuder Preference

Record, suggests that the main reason for the invalidity of the personality inventory probably is a fear of others, based on a feeling of personal insecurity. Subjects who gave the more valid written self-descriptions were more direct, more outspoken, and less dependent on others.

REFERENCES

1. **Kuder, Frederic**
   1944. The Kuder Preference Record. Science Res. Associates. Chicago
2. **Piotrowski, Z. A.**
   1937. The Methodological Aspects of the Rorschach Method. Kwart. Psychol. **9**: 29–41. Also 1936. **Rorschach, H.** Res. Exch. **1**: 23–28
3. **Piotrowski, Z. A.**
   1942. A Comparative Table of the Main Rorschach Symbols. Psychiat. Quart. **16**: 30–37.
4. **Piotrowski, Z. A.**
   1942. On the Rorschach Method of Personality Analysis. Psychiat. Quart **16**: 480–490.
5. **Rorschach, H.**
   1921. Psychodiagnostics: A Diagnostic Test Based on Perception. H. Huber Berne (Switzerland). 1st ed.
   1942. English ed.: 226.

DISCUSSION OF THE PAPERS

Dr. F. L. Wells: In our attitude toward clinical psychometrics, there is no more encouraging feature than the growth of interest in patterns. In the earliest application of the Stanford-Binet series for psychotic cases, it was common to observe how the affective and schizophrenic symptom-complexes reflected themselves in particular responses. Anyone familiar with both could make a good guess at how and where: in digit span; in the comprehension questions; in the responses to pictures; the absurd sentences, etc. But this type of examination does not lend itself well to study by patterns The early work of Pintner and Paterson did so in organization, but its content was, from this standpoint, limited. The general chairman of this Conference had a conspicuous part in laying foundations for this approach, in the ground-breaking study of adult intelligence published in collaboration with Weisenburg and McBride, about ten years ago. A year later, some highly significant contributions were published by Jastak and others of the Delaware group. The importance of this approach was well established, but these investigators had to do all their work with instruments picked up from here and there, like the Binet scales, Formboards, and Porteus Mazes, by no means designed for such coordination. The Wechsler-Bellevue scale was the first to give us such a series of instruments, and though its significance in this respect was not at once understood, there can be no doubt that the present developments have sprung essentially from Wechsler's work. Other useful measures have been added, but their usefulness rests in organization of the Bellevue type; that is, series of tests strong enough to stand alone, and thus to give mutual support in establishing the psychometric pattern.

My general experience with the inventory type of procedure permits only the limited testimonial that the Scotchman gave his departing clerk, "regarding his honesty I can say nothing, because I never trusted him" It seems clear, however, that, in the stress of present circumstances, advances have been made, both in technique of construction and manner of evaluation, that will entitle them to

wide use, as they become generally available and understood. From some standpoints, it is unfortunate that their immediate function has made them so largely catalogues of maladjustive symptoms Only the Kuder really stands out in the contrary direction, and it is pleasing to note that personnel officers have taken ready note of its potentialities. Apropos of diagnostic significances, the sounder use of these procedures is probably as indicated quite generally in these studies, along the single gradient of adjustment or maladjustment, leaving qualitation to more direct observations One may note that the use of cut-off points emphasizes a continuity in this gradient, which probably does not obtain in psychotic conditions. The word-choice (essentially a closed end form of the old free association test) is a very ingenious device, which has a number of possibilities in connection with the concepts that developed about the older form of the test. I would raise the question if the large number of false positives reported for the Cornell Selectee Index by Drs. Weider and Wechsler does not embody some elements of malingering. It would be suggested by their TABLE 3 data as compared with TABLE 2. One may note Harris's "meticulous truthfulness" comment on this point; possibly, an unconscious malingering. In this connection, when one asks about a person's liking for occupations, one must distinguish whether the answer represents what the man does like, or rather what he would like to like, a sort of censorship by the superego. Dr. Piotrowski's paper has relevance to this point I understand that this often complicates the interpretation of thematic productions. Lt. Comdr. Harris, in particular, suggests the value of these procedures in gauging changes of adjustment level over periods of weeks or months.

I am not sure that the term, "Diagnostic Testing," to distinguish the configurational approach, is an altogether happy one. There comes to mind a remark of Adolf Meyer's at a psychiatric conference; we understand this case, we don't need any diagnosis. Diagnosis does not add to information about a case, it merely condenses and may distort it. It is always but a means, and a means to classification rather than management, where nosologic boundaries are as uncertain as they are in neuropsychiatry. It has been remarked, elsewhere, that psychotic diagnosis is itself not without projective features Thus, I look a little askance at any psychometric device that heads directly into the diagnostic entities of this field. The proper function of these devices is more precisely to delineate symptoms which may have one of several meanings What is the meaning of a low digit span forward we judge by its relation to digit span backward, to vocabulary range, to arithmetical reasoning, to story memory, not to say data extrinsic to psychometrics. And that leads us into an understanding of the case in terms of what to do about it, which is more than capable of bypassing the diagnostic label altogether.

This is necessarily so, in the different field where I have been occupied, since there we have very few such labels that are not less than useless—like introvert-extrovert, for example. One's accumulating experience in this field tends to crystallize in terms of trait-categories, as they might be Thurstone's Primary Factors or Sheldon's temperamental features Present attempts at diagnostic concepts in this field would be intolerable over-simplifications Psychopathology seems to tolerate them better, but an understanding of processes involved, at which Rapaport aims in his current volume, is more important than any tie-up possible with present diagnostic labels

Dr. Wladkowsky mentions disparities of verbal and performance scores in psychotic boys, but a crucial feature here is its relation to the nature of the psychopathy "Conduct" problems have been said to be higher in performance; "personality" problems, so-called, higher in verbal tests; which might yield a very interesting tie-up with the somatotonic and cerebrotonic types of Sheldon.

Higher digits backward than forward scores probably have various sources, but a long memory calls to mind the analogous observation of Franz, who reported a clinical picture showing shorter choice reaction times than simple ones. He attributed this to the greater challenge of the more difficult process; something similar could be operative here.

Dr. Wladkowsky's point is well taken, that tests are not projective or non-projective, but differ in the degree of their projective attributes. In multiple choice

intelligence tests, these are practically non-existent; in Binet tests, they are present but slight; in Rorschach or TAT, they are extreme. The good clinician keeps on the alert for projective features in intelligence tests, as he would for intellectual levels disclosed in Rorschach responses.

It is creditable that so much of this configurational work has been accomplished with the instruments at hand, but now that we are outgrowing the concept of the I. Q. and other global indices of this nature, we could well look towards a refinement of the procedures which sprang directly from this earlier tradition. It seems to me that each and all of the Bellevue sub-tests would stand some strengthening for the role that they are here assuming. The arithmetical portion, particularly, could be amplified, with more of a separation between computational and reasoning functions; a finer gradation of responses would be helpful at various points Dr. Schafer develops this matter in some detail. If this configurational viewpoint is to develop as it should, it needs procedures more refined than those currently available for most clinical work.

From all of this, it emerges that, with verbal formulation in clinical psychometrics, there is really no escape from oversimplifying. There are few mental disciplines where the printed word gives so little in proportion to field experience; and in order to be useful for didactic purposes, this experience must have breadth as well as depth. *De mortuis*, indeed. But it is well to remember the errors to which even Rorschach was subjected through an experience so largely limited to the pathological. One whose experience is with the neurotic will find many gaps in the ideas of one who has looked only at the psychotic Those who deal with juvenile behavior problems will get different ideas from either about the meanings of vocabulary, or block designs, or mazes. Every test is a projective test, indeed —on the one who gives it.

Roy Schafer (*Menninger Clinic, Topeka, Kansas*): Drs. Wechsler and Wladkowsky say that the Bellevue Scale is first of all an *intelligence* test, capable of simple and successful diagnostic application. I submit that this distinction is not a valid one: intelligence functioning, as seen in clinical work, is so much shaped and warped, impaired or sharpened, by maladjustment features and different kinds of personality-organization that it becomes merely another area of expression of the person himself, and does not stand off by itself as one of his possessions. A dynamic view of intelligence is indispensable.

Dr. Wechsler appears to maintain that successful diagnostic work with his test is largely "inspired guesswork." Granted that, in the early application of any test, much intuitiveness without conscious formulation of concepts and relationships does come into play, it is the responsibility of clinical psychologists to organize their experience and material by systematic research, in order to render "objective," quantifiable (as far as possible), and communicable, their diagnostic clues Diagnostic work can be taken out of the realm of "inspired guesswork." This paper and the investigation from which it is derived represent an attempt in this direction.

Dr. Wechsler states that the cases described in the two papers on the Bellevue Scale were on the order of "good stories"; that is, instances where, somehow, it all became clear. But they have more than anecdotal interest: these cases, in which the dynamics of the illness and their effect upon intelligence functioning become more or less clear, are of crucial significance for the practice and theory of clinical testing. For theory, they indicate meaningful relationships which the psychologist will eagerly seize upon, for their value in helping him develop an understanding of the tool he is using, of the functions he is measuring and of the general theory of the test as a whole; for practice, the clinician long remembers such cases and looks forward to more, because it is around such clear cases that his "experience" becomes organized: from them, leads for systematic investigation can be found. The Bellevue Scale is a valuable clinical-experimental tool, one which can serve to demonstrate fundamentally important psychological relationships

NON-PROJECTIVE PERSONALITY TESTS

PART IV

THEORY

structured refers to a clear and unequivocal meaning, the term *unstructured* refers to an unclear equivocal material to which the subject gives meaning. To put it simply, in projective tests, the subject does not know what reaction is expected of him; in non-projective tests, he knows clearly. This simple statement is, however, misleading. To be correct, it would need many modifications. For instance, one would have to state that, in the projective Rorschach Test, there is no unequivocal, socially accepted, logically exclusive and 1:1 verifiable reaction; while, in a non-projective test, such as an intelligence test item asking, "What is the capital of Italy?", there is a definite, expected reaction which can be verified, and concerning which there is a definite and logically indisputable agreement. In other words, one could say that the non-projective tests measure the subject's reaction to, knowledge of, and compliance with, general agreement. The projective tests do not have such standardization. But even this formulation has weaknesses in its reference to both the projective and the non-projective tests. First of all, the projective tests also establish "popular" trends and even expected norms, the differentiations from which constitute diagnostic indications. Second, the non-projective tests do not *all* ask for statements concerning which verifiability, social agreement, and logical necessity are all present. For instance, such a question as that in Bell's Inventory, "Do you feel tired most of the time?" is one, the verifiability of which is of a different order than the verifiability of the capital of Italy. Social agreement on the response can be had only with great difficulty, would be of questionable value, and no logical necessity would be attached to it.

There are other difficulties in our simple formulation of the difference between projective and non-projective tests. The subject may know that, on an intelligence test, he is expected to give a definite, factually correct answer, or that, on a questionnaire, he is expected to give a statement of fact; but on the projective test also, the average subject, in general, assumes the same. Furthermore, in both kinds of test, the individual response (concerning the *meaning* of which the subject may have some very definite ideas) derives its *significance* for the examiner, not from itself, but rather from its statistical relationship to other responses of the subject and to response patterns of the general population. If we then exclude those tests which are mere questionnaires, replacing other modes of quest for information of which the subject is consciously in possession, the differences between projective and non-projective tests of personality appear to dwindle.

At this point, it would seem that this paper, which set out to state the principles underlying non-projective techniques of personality appraisal, has become a funeral march to the tune of which the non-projective techniques are to be buried in the mass grave of projective procedures So let us state explicitly that the original definition—that projective tests deal with unstructured, while non-projective tests deal with structured, material—does hold true and has far-reaching consequences. In the projective test, the subject organizes or structures an unstructured material and, in this structuring, reveals his own psychological structure. In the non-projective tests, the responses are usually not those which bring about structuring; but the totality of responses, when compared with the trend of responses of large populations, proves to have a structure also reflecting the subject's personality structure. It is as though, in the projective tests, the conclusions are drawn from the manifestations of the *active functioning* of the subject, while in the non-projective tests, the conclusions are drawn from the patterns of his *conventional* responses To be specific: In a test like the Rorschach, we see perceptual and associative processes at work in new creation, and we must infer their nature from their work. In a test like Similarities in the Bellevue Scale, we see the subject's verbal concepts; in Comprehension, his formal judgment; both crystallized, both quasi-stationary, and both born in the course of a long history of active, associative, creative work. They are so stationary that their function character frequently eludes us altogether Both the active functioning of a personality and its crystallizations in the inter-relationships of the quasi-stationary structures, such as judgment and verbal concepts, are revealing of the personality. The attack of the projective and non-projective techniques of personality appraisal differ from each other, in so far as the projective attack attempts to tap the active principle of the personality, while the non-projective attack is aimed at the personality's quasi-stationary structures. *Thus, the first assumption underlying non-projective techniques is that there are, on the one hand, active, non-stationary psychological functions; and, on the other hand, quasi-stationary functions*

II

Let us now consider more closely these quasi-stationary psychological functions we have referred to. Under different conditions, our motives and desires may choose different pathways and may be modified and altered, and take devious courses toward their goals; and thus they

will appear as non-stationary psychological functions. But, if we want to account for any interpersonal understanding and agreement, we must assume stationary structures in our psychological life, or within that segment of it to which we refer as intellect. Now, we may look upon intellect as a storehouse of static assets, or static liabilities, or we may look upon it as comprising various abilities that are being used by our motives and wishes whenever expedient. Thus, for instance, one may think of verbal concepts simply as static assets possessed by an individual, or as tools once acquired and used when expedient. However, if we realize that we create concepts steadily (the scientist does it in all of his moves, and so does the man in the street in many of his moves), we shall find it nonsensical to consider our crystallized verbal concepts as something different from our active concept-creating activities. Rather, we shall see verbal concepts as sediments, crystallized quasi-stationary forms of these active concept-creating functions. The transition between the active, non-stationary functions and the quasi-stationary ones appears to be fluid and continuous. In other words, the more the conditions calling for use of concepts can be met by verbal concepts already formed and crystallized (we call these acquired), the more we are entitled to speak about a quasi-stationary function being at work. Underlying it, however, there appears to be a function with few stationary characteristics—a function which is flexible and modifiable, and, thus, is related to the mode of functioning of man's motives and desires. Obviously, we must assume that the quasi-stationary function, represented by verbal concepts, is not the only quasi-stationary function of our psychological makeup.

Let us here turn from our consideration of these quasi-stationary functions, and focus again on their relationship to personality appraisal If it is assumed that these quasi-stationary functions have been created by active non-stationary functions, then the wealth, stability, accuracy, etc., of these quasi-stationary functions, as demonstrated in their test achievements, must reflect the strength and the development of the active functions, as well as the encroachments of maladjustment upon them. If, then, the wealth and stability of these quasi-stationary functions can be compared with each other in terms of their test achievements, we should obtain a picture of the relationship of the wealth and strength of the different active functions underlying these "tools." In our work on personality evaluation by means of intelligence and concept formation tests, we found that the relationships thus assessed are characteristic for different types of personality de-

velopment. Therefore, from the intercomparison of the performance or achievement of the quasi-stationary functions, inferences can be drawn as to the active functions underlying them; and from the inferred relationships of underlying functions, inferences can be drawn as to the type of personality, or type of encroachment upon personality development, characteristic of a subject.

However, this is easier said than done. An achievement, usually, does not imply a single quasi-stationary function, but many of them; and from this fact, certain consequences ensue. (1) It is advisable to consider these quasi-stationary functions as aspects of, that is, as our mode of looking upon, our psychological functioning, rather than as something that has its own independent, actual existence. Thus, in our work, we found that memory, concept formation, attention, concentration, anticipation are some of the quasi-stationary functions we had to postulate as aspects of our psychological functioning. (The precise meaning of these concepts, as used here, is discussed in our volume, "Diagnostic Psychological Testing.") When, in the Similarities test on the Bellevue Scale, the question is asked: "In what way are a dog and a lion similar?", there can be no doubt that the subject must first of all *"attend"* to the question itself; must make a correct *anticipation*, in order not to come out with a statement of the opposites, or with descriptions of the dog and of the lion, instead of a common denominator of the two; that he implicitly performs a *memory function* in correctly invoking the characteristics of dogs and lions; yet, though attending, anticipating, and remembering are all involved, the preponderant function remains one of verbal concept formation. Such clear preponderance, however, is by no means the usual case. (2) Another consequence is that it is difficult, therefore, to determine how many, and which, are these quasi-stationary functions. In working with the Bellevue Scale, we had to assume the existence of such quasi-stationary functions as concept formation, memory, the triad of attention-concentration-anticipation, visual organization, and visual motor coordination.

Whether or not the dividing line between these quasi-stationary functions and the non-stationary functions, here discussed, follows the dividing line of *ego* and *id* in the psychoanalytic sense, is a moot question. It is quite possible that the line dividing the *ego* and *id*, and that dividing the conscious and not conscious parts of the *ego*, will be drawn more sharply, if and when the relationship between the basic motivating forces, the non-stationary functions, and the quasi-stationary functions is better understood. This does not imply that projec-

tive tests explore the *id* directly. It is clear only that these quasi-stationary functions have much to do with the *ego*, and exploring them is an important part of what is called *ego* psychology. It is the issue of the exploration of the psychology of the thought processes. *Thus, our second assumption underlying non-projective techniques of personality appraisal is that non-projective tests of personality are to be so constructed that their parts correspond to quasi-stationary psychological functions; and they are to be so construed that they allow for comparison of these quasi-stationary functions with each other, within the individual and against the norm for this relationship, in the total population; because, only out of the relative strength of these functions shall we perceive the variants of developmental conditions, characteristic for the personality.* A corollary of this assumption is that non-projective tests of personality are to be based on definite assumptions as to the functions underlying responses to their different parts. Therefore, non-projective personality tests are a tool, not only of diagnosis, but also of exploration of the psychology of thinking.

III

My discussion, thus far, has referred to, and been based on, intelligence and concept-formation tests, used as non-projective tests of personality. Even admitting the validity of these considerations and assumptions for these tests, it might still be objected that we have not demonstrated, or even suggested, why these assumptions should be valid for all non-projective tests of personality. The objection is sustained, and no claim is made here as to the general validity of these considerations and assumptions. Nevertheless, it will be worth while to examine whether or not we can cite further considerations making it advisable to keep these assumptions seriously in mind as applicable to other non-projective tests of personality.

It might be argued that, in order to construct a successful non-projective test of personality, it is not necessary to have any kind of theory or assumptions as to quasi-stationary or other functions underlying test responses. It might be argued that, by giving a sufficiently large and varied set of questions to a sufficiently wide normal population, which has sufficiently large and well-defined neurotic and psychotic subgroups[4] of all important varieties, one will be able to find sets of questions, the responses to which will reliably differentiate all the major groups, as well as all the subgroups from each other. Such an argument I could support with a series of data concerning such successes.

All these data, however, have one or more of the following three features in common: (1) The application of the tests was successful in limited groups, which in age, background, and education, were highly similar to the standardization group. (2) The differentiations afforded by the tests were few and gross, that is to say, they sought to segregate successful or non-successful students; they stated, "he is all right," or "there is something wrong with him"; they attempted to differentiate psychotics, neurotics, and normals; and, perhaps, attempted a further bi-division, in the psychotic range, by differentiating affective and schizophrenic psychoses. (3) The tests pertained to one very specific aspect of the personality (as, for instance, vocational interest), and the successful tests among these used homogeneous sets of questions.

The data, which these tests thus afforded to support the contention that statistical procedures alone are sufficient, really pull the last bit of ground out from under this contention, and prove that, in the long run, this statistically-safeguarded dream of machine-like personality assessment is a hopeless illusion. What chemist would agree to use chemical tests which will detect acidity only if it occurs under certain definite conditions, and particularly only under the same conditions under which it has been seen already in the past? A test is testing only if, *under new conditions*, it can detect the old and known elements (in chemistry) or relationships (in chemistry, as well as in psychology). Conditions are always new, even in our inorganic world; more so, in our organic world; and, most of all, in our psychological world. Tests are useful and of general validity only if they are not limited to specific conditions, but are, in themselves, able to cope with variations of conditions, and either lead to a detection of the entity or relationship that is being sought, or indicate specifically the other testing procedures to be applied under the changed conditions.

The two concepts of testing here contrasted are not merely the contrast of two testing procedures; they are expressions of two diametrically-opposed views of science and of the world. They are expressions of the contrast between mechanical, statistical, atomistic, pragmatic views and functional, dynamic, organismic views. For the atomistic view, science is a matter of probabilities, theories are illusions, and, instead of theories, correlation coefficients assume the role of the demigod. For the organismic view, science is a search for functional relationships and laws; the keystones of science are hypotheses, in terms of which data and observations can be systematized; and statistics, the most important crucial testing tool of the validity of these hypotheses,

remains merely a *tool*. This view of science does not expect statistics to take care of thinking, or to reveal relationships. It does not trust statistics more than human experience and intuition. It relies on statistics only to check, test, systematize, verify, cleanse, and build into communicable form what is given in human experience.

Let us return to the three conditions which we found operating in successful non-projective tests built upon the purely statistical approach: (1) that the groups tested must be limited and similar to the standardization groups; or (2) that the differentiations are few and gross; or (3) that the tests utilize homogeneous sets of questions. On the first two points, we may ask, "How is it that statistically-designed procedures will work, to some extent, even for limited groups, and yield even gross differentiations?" Cultural patterning, and its tremendous impact upon personality-structure, supplies the answer. The situation is somewhat different, as regards the third point, the homogeneous structure of the tests. To illustrate: In a test like Bell's Adjustment Inventory, one finds a question like, "Has it been necessary for you to have frequent medical attention?" together with a question like, "Do you get angry easily?" One question asks for a statement of fact, the other asks for a subjective appraisal of one's subjective experience, both as to degree and as to frequency. Inventories, questionnaires, and other non-projective tests of personality are generally inclined to mix indiscriminately questions pertaining to different levels of the personality, and the response to lack of these will, therefore, be based on very different quasi-stationary functions of the individual In a test like the Vocational Interest Test of Strong, we do not find such mixtures. We find, rather, several, well-defined, structured sets of questions, all essentially similar to each other, asking for preferences in vocations, in amusements, in activity, in people, in styles of living, in school subjects; though, regrettably, we also find a group of subjective, self-rating items. The explanation of the efficacy of this test is based, in our appraisal, on this consistent (for the most part) uniformity of the questions. The questions pertain to interests. Interests, also, are apparently quasi-stationary functions of personality, although we know as yet very little about their dependence upon total character structure. Thorough personality studies, incorporating this very successful Strong test, may reveal to us those non-stationary functions which underlie interest-formation. They may reveal to us what the variants of the quasi-stationary functions, called interests, are, and what is their systematic place in psychological functioning. It is quite

possible that, sooner or later, if we explore many other such quasi-stationary functions, we shall have ways to reconstruct the relationship between these and those underlying interest-formation. From these relationships, we shall obtain more clear-cut pictures as to the quasi-stationary structure of the individual personality. At this point, non-projective testing of personality will be as efficacious as, or more efficacious than, projective testing of personality, as it is at present. There is little doubt in the author's mind that, for educational and vocational advice, for counseling, and for industrial testing, the significance of non-projective tests of personality is greater than one can perceive for the projective personality tests extant. How projective and non-projective tests will complement each other, is a question to be answered in the future, on the basis of factual findings. *Our third assumption underlying intelligence and concept-formation tests used as non-projective tests of personality* (and which recommends itself to be applied to *all* non-projective tests of personality appraisal), *is the necessity of having a definite concept and theory of the quasi-stationary functions underlying the reactions and achievements on these tests.* A corollary to this assumption is that it is desirable that the questions or problems of these tests be homogeneous, or that the test consist of internally homogeneous item-groups. In this way, they will each tap different quasi-stationary functions, or different variants of a quasi-stationary function, and render them subject to comparison. Out of these comparisons, characteristic patterns of quasi-stationary functions can be derived, and the relationship of these patterns to specific types of personality developments and maladjustment may be inferred.

IV

Let us consider, finally, the nature of quasi-stationary functions, and their relationship to the non-stationary functions. We have already stated that it is both expedient and in accord with our experience and analyzed data to conceive of these quasi-stationary functions as crystallized sediments of non-stationary functions, intimately related to motivating forces of the personality. How should we conceive of the relationship of the non-stationary and the quasi-stationary functions to these motivating forces? It will not be possible for me to adduce evidence within the limits of this paper, to substantiate the view of this relationship to which we adhere in our studies. Therefore, I shall make only a brief statement of it.

Drives or motivating forces of an individual may undergo, in the course of his development, different vicissitudes. These vicissitudes will mold what we refer to as his personality. They may lead to the attitude, well characterized by the motto of the proverbial three monkeys, "I hear no evil; see no evil; speak no evil." It has become a custom to refer to the fate of drives resulting in such a motto as "repression" and "inhibition." Since everything in the world of perception and action may bring danger, the field of action, as well as that of perceptual intake (attention), becomes limited. Here we see, then, a vicissitude of the drives expressing itself in quasi-stationary forms. The quasi-stationary functions underlying information and the function of attention become seriously limited. Another vicissitude of drives is seen when the threatening danger is responded to by alertness keyed to the highest pitch and is ever-present It is the custom to refer to the results of this vicissitude as "obsessiveness" or "compulsive meticulousness." The function of attention becomes overemphasized; the quasi-stationary function underlying information works intensively and extensively; verbal concepts become sharp and rigid.

These examples are given to illustrate the thesis that the quasi-stationary functions discussed here really reveal *the mode of control of drives and impulses*, of the fundamental motives of all psychological life. We, as yet, have few, if any, reliably direct measures of the native strength of these drives and impulses, and what little we know qualitatively about personality structure refers mainly to their mode of control. Thus, one can justifiably state that, for the time being, we distinguish one personality from another by reference to these different types of mode of control. Non-projective tests of personality dealing with the quasi-stationary functions discussed above make these modes of control (or rather the crystallized results of such modes of control) palpable and testable.

PROBLEMS OF PERFORMANCE ANALYSIS IN THE STUDY OF PERSONALITY*

By Martin Scheerer

College of the City of New York, N. Y.

I. THEORETICAL PREMISES

Theoretical considerations may mislead, if they are not linked with adequate experimentation. Experimentation, however, may misinform, if not linked with adequate theory. It is, therefore, the purpose of this paper to attempt a theoretical discussion of performance analysis, and, at the same time, to connect its propositions with concrete experimental problems and data.

What is a performance? I propose to define performance "holistically": every performance is the expression of the organism's activity as a whole, and not a segmental response to a specific stimulus.

Recent views of the organism as a dynamic whole have been couched in various formulations. I may remind you of Jenning's statement that "the organism is a process," of L. K. Frank's[1] reference to the field concept which "leads to a reformulation of the idea of 'parts' and 'whole' and of the problem of organization," or of A. Angyal's postulate that "the study of living beings in general and of the person specifically should be a study of the organismic total process, the study of the processes of living."[2] Can the nature of this "living whole" be wrested from speculative generality and be made experimentally tangible? The answer seems to lie in the studies that have led to this very conception and have already proven their fruitfulness, as in the work of W. Stern,[3] G. W. Allport,[4] the Gestalt psychologists, and K. Goldstein.[5] Goldstein's conclusion that human behavior is a process of "self-actualization," of "coming to terms with the surroundings," is particularly supported by his clinical findings that symptoms are "answers"

---

* This is a preliminary report resulting from a larger study to be published in book-form with the support of a grant from the American Philosophical Society.

[1] Introduction to the conference on psychosomatic disturbances in relation to personnel selection Ann N Y. Acad Sci. **44** (6): 541–549. 1943.

[2] Basic sources of human motivation. Trans. N. Y. Acad. Sci. **6** (2): 42–57. 1943.

[3] General Psychology from a Personalistic Standpoint. The Macmillan Company, New York. 1938.

[4] Personality; a Psychological Interpretation. Henry Holt and Co., New York. 1937.

[5] The Organism. American Book Co. New York. 1939; Human Nature in the Light of Psychopathology. Harvard University Press, Cambridge, Mass. 1940.

of the impaired organism in trying to cope with the environmental demands to the best of its capacities.

All of these views, also the most recent developments in psychosomatic medicine, seem to have further in common the premise that "even the structural features of the organism such as the morphological pattern of the heart, or the lungs, have their full meaning and significance only in relationship to the function that they are carrying out within the total organismic process."[6] In this context, it is of more than historical interest that, since 1917, J. S. Haldane has been accumulating physiological evidence for the claim that "structure and functional relation cannot be separated . . . since structure expresses the maintenance of function and function the maintenance of structure. The problem before the biologist is to discover how these apparently isolated and unrelated phenomena express aspects of inherent coordinated activity."[7]

If we accept the position implied in these formulations, then the definition of performance would not be the formula: stimulus-response plus the sum total of any other processes and their interaction, but rather the formula: in a given situation, the entire organism deals with that situation by structuring it in terms of a definite figure-ground pattern and acting with reference to that pattern.[8] That pattern will tend to become salient which has *functional* relevance for the organism, that is to say, which has a meaning adequate to the person's natural and acquired dispositional equipment, and to his current activity.[9] More precisely, the pattern that will form is a figure-ground *experience* which is part and parcel of a *behavioral* figure-ground organization; it is *one* action process of dealing with the situation. A performance, then, is "adjustive" or "adaptive" behavior, to use G. W. Allport's terms,[10] and, as such, it is a concrete phase in the process of self-realization which is not merely reactive, but spontaneous and creative, as well.

Following up this premise, we can expect that every performance will be characteristic of the individual's unique way of coming-to-terms with the surrounding field. Following William Stern's suggestions, however, we should expect that the whole person will manifest his in-

[6] Cf. [3] l. c.
[7] Organism and Environment as Illustrated by the Physiology of Breathing. Yale University Press. New Haven, Conn. 1917; The Philosophical Basis of Biology: 22-31. Hodder & Stoughton and Co. London. 1931.
[8] Cf. [4] Also Köhler, W. Gestaltpsychology. Liveright Publishing Corp. New York. 1928. Koffka, K. Principles of Gestaltpsychology. Harcourt Brace and Co. New York. 1935. Lewin, K. Dynamic Theory of Personality. McGraw-Hill Publishing Co., Inc. New York. 1935.
[9] Cf. [5] Also Freeman, E. Social Psychology: 152. Henry Holt and Co. New York. 1936.
[10] Cf. [4] l. c.

dividual characteristics, not to the same degree, but to different degrees in different activities. For example, he would express himself with varied uniqueness, in handwriting, in projective test or problem-solving situations. With this qualification, not only the thought processes should bear the stamp of the person's organizational matrix, but perceptions, feelings, and motor-acts as well On second thought, these apparently different processes may actually be one unitary performance in which no true separation exists, but, instead, one definite pattern, in which emoting, thinking, and perceiving articulate in a configurated dynamic relation to each other. If all this be true, then the terms, perception, emotion, and thinking, in their separate applications, would be only conventional abstractions which the psychologist is forced to make, in order to maintain the control of otherwise boundlessly merging and numerically overwhelming variables.

It is on these grounds that the scientific conscience of many investigators justly opposes the postulates of organismic psychology and adheres to the traditional abstractions. If we want to overcome this *impasse,* we should perhaps look for other categories of behavior and for such experimental techniques which permit us to capture the essential holistic relatedness in the performance of the person under scrutiny. Of course, I am not as yet in the position to present any elaboration of such categories and techniques. I may, however, submit a few relevant methodological considerations which might help to achieve this end, and point out some pertinent experimental developments. Let me first turn to the problem of techniques.

II. PRINCIPLES OF TECHNIQUE

Here I should like to single out two aspects of adjustive behavior: first, performances in which the cognitive aspect is in the foreground; second, performances in which the "perceptual-motor" aspect is in the foreground.

The Cognitive Aspect

Notwithstanding the great accomplishments in intelligence testing, we have less information about the factors that underlie the profile of scores than we would like to have. The outstanding attempts in this direction are factor analysis and that of the grouping of scores. These attempts have been more and more supplemented by a *qualitative* study of the subject's mode of approach, as is now done at the Menninger Clinic and originally had been emphasized by Goldstein,

planation is called for. Because of their exemplary value, may I refer to the drastic findings in certain brain injuries?[18, 19] Some patients cannot correctly read or recognize visual material in *short* exposure, but appear quite successful without time restrictions. Here, however, they also fail, if prevented from moving their head or hand during attempted recognition. Exploration reveals a visual agnosia owing to which no forms are seen, but only a contrast between spongy masses of different color or brightness. While step-wise tracing along the contrast border, the patients make accompanying head or hand movements and infer, thereby, from their *kinesthetic* experience the form of the object. This, naturally, takes time. Here, the probing into the qualitative conditions of failure led to the discovery of the conditions for success, of its detour character and to a diagnostic understanding of the patient's changed capacities.

Evidently, the study of performance has to determine the psychological conditions of both failure and success. Without this, no consistent scientific theory of mental functions is possible, let alone a conception of how they are organized in a given personality. I, therefore, propose that the definition of success or failure in a test should be made dependent on the following principles:

1. We need a psychological analysis of the task which the test item presents. This demands a "phenomenological" and experimental identification of those processes which are *requisite* to the solution. If it is found that the same solution can be achieved by various alternative processes, either the task must be altered so that it permits only one type of solution, or the subject must be prevented from using the alternative procedure. Aside from certain test constructions, to which I shall refer later, this proposal somewhat parallels present methods of "job analysis."

2. With this postulate fulfilled, the scoring would no longer be *external* to the procedure, but would be a scoring *of* the procedure. With that, we would be in a position to inquire into the capacity or ability which is the functional precondition for the processes leading to the solution. In brief, achievement and process would now represent one measurable and unitary dimension of performance.

3. In determining the psychological requirements for task solution, we must also consider the hierarchic order of functional levels, the dif-

[18] **Gelb, A., & K. Goldstein.** Zur Psychologie des optischen Wahrnehmungs- und Erkennungsvorganges. Ztschr. f. d. Neurol. u. Psychiat. 41: 1. 1918. Also in **Ellis, W. D.** A Source Book of Gestalt Psychology: 315. Harcourt Brace and Co. New York. 1938.

[19] **Ranschburg, P., & E. Schill.** Ueber Alexie und Agnosie. Ztsch. f. d. gesamte Neurol. u. Psychiat. 1932.

ferent capacity stages on which a subject may operate in reaching the same overt result. In children, these stages are developmentally conditioned, while in the adult it is motivation, attitude, or set that will determine on which level he copes with a task. It has been shown that the child who is in the stage of using number words in concrete series, who can count off 5 fingers or 5 apples in a row, may not yet understand '5' as a collective sign for the combination of 5 objects, or has no simultaneous grasp of 5 arbitrarily grouped elements as a whole unit of '5' Moreover, this number mastery of perceptually grouped units is yet to be followed by the conceptual stage where operations with numbers, as such, in the abstract can be carried out.[20,21] We, therefore, have to decide for which genetic level we are testing the number function in the child, and we must introduce the necessary task of transposition if we test for a higher level.

In adult testing, we are confronted with analogous problems This has been convincingly demonstrated by F. L. Wells,[22] who varied experimentally the "ball and field" test to determine "the plan of search at various levels of abstraction." On the basis of her investigations, E Heidbreder[23] recently expounded the view that we are dealing with a hierarchy of cognitive organization in human beings. According to Heidbreder, we have a basic level of concrete perception of objects, and a conceptual level which stands in an integrating relation to the first Her empirical evidence, which parallels the results with sorting tests discussed by Goldstein and Scheerer,[24] supports the proposal to construct tests in such a fashion that we can determine, in advance, on which level the solution shall occur. This becomes crucial, in testing for impairment of abstract behavior. We know, from H. Babcock's[25] work, that the efficiency index shows a relative constancy of the vocabulary score as opposed to the non-vocabulary score. Her explanation that the involved symbol functions are "older habits," and, therefore, not so affected as others, has been challenged by various investigators. G. K. Yacorzynski[26] has criticized the Babcock law on the grounds that this test "makes use of only the end results in the definition of

[20] **Werner, H.** Comparative Psychology of Mental Development Harper and Brothers. New York. 1940. **Strauss, A.** Problems and method of functional analysis in mentally deficient children. J Abn. Soc Psychol. **34** (1): 37–62. 1939.
[21] **Riess, A.** An analysis of children's number responses. Harvard Educ Rev. 149–162. March, 1943
[22] J. Gen. Psychol. 21: 163–185. 1939.
[23] Toward a dynamic psychology of cognition. Psychol. Rev. **52** (1): 1–22. 1945.
[24] Cf. [11].
[25] An experiment in the measurement of mental deterioration. Arch. Psychol. **117**. 1930.
[26] An evaluation of the postulates underlying the Babcock deterioration test. Psychol. Rev. **48** (3): 261–267. 1941.

words and therefore superficially it appears as if the vocabulary remains unaffected by the general deterioration." In support, he cites H. M. Capp's[27] findings that a word which could be defined in at least 5 different ways was given only two meanings by the most deteriorated patients, against 4 to 5 by the others. But the deteriorated is not only less able to shift from one meaning of a word to any other; he also gives fewer definitions, because he grasps only the *simpler* meanings of that word. This is substantiated by the higher correlation of deterioration with the scores on written tests than with those on oral tests. On the oral test (*e.g.*, Binet vocabulary), many solutions were possible, depending on the level of ability at which the subject was functioning. Thus, if he was only capable of giving a simple solution, he was credited. On the written test, the correct answers were predetermined and allowed the subject no alternative in obtaining the solution. If this was beyond the scope of the subject, he would fail the item, since he could not substitute his own solution on a simpler level. We agree with Yacorzynski's interpretation that, if correct responses can be obtained by different procedures, the "easier methods of reaching the same end results are left to the organism, even if the more difficult conceptual organizations are no longer available."

As here suggested, current research is placing more emphasis on the qualitative aspects of the patient's definitions and his procedure on verbal tests of *increasing* difficulty. S. B. Ackelsberg[28] studied 50 senile dementia patients of equivalent educational background and age, with Capp's tests and 12 homographs. In substance, she corroborated his results, particularly on the synonyms and antonyms. "The vocabulary functioning does not remain constant; on the contrary these tests . . . have given measures of deterioration." S. E. Cleveland and D. W. Dysinger[29] applied the Goldstein-Scheerer sorting and the Bellevue-Wechsler tests to senile psychotics, and found that these were unable to sort objects on a conceptual basis, but passed many of the verbal items on an *apparently* abstract level. The authors conclude "that this appears due to the fact that the patient may use what seems to be an abstract verbal concept with much more restricted meaning." These studies need not detract from the clinically diagnostic value of a discrepancy between high verbal, and low non-verbal, scores.[30] They do, however, argue

[27] Vocabulary changes in mental deterioration. Arch. Psychol. **242.** 1935.
[28] Vocabulary and mental deterioration in senile psychosis. J. Abn. Soc. Psychol. **39** (4): 393–406. 1944.
[29] Mental deterioration in senile psychosis. J. Abn. Soc. Psychol. **39** (3): 368–372. 1944.
[30] Cf. Shipley, W. C. A self-administering scale for measuring intellectual impairment and deterioration. J. Psychol. **9**: 371. 1940.

for the proposition that the abstract function represents a unitary organizational level. As far as there may be several such levels, impairment of abstract function might vary in degree, but will to that degree affect methods of reasoning. Space does not permit presentation of further evidence on this point, as it has accumulated in abnormal and developmental psychology—in the latter, also, through the use of ingenious conditioning methods.[31]

4. The study of varying 'levels' and of different 'avenues' of procedure in the task solution is not alone important for the experimental control of these variables. Their identification may also serve as a clue for personality differences, because the subject may show a preference for, or a dependence on, a particular level or avenue of approach. E. Hanfman,[32] for example, giving the Vigotski test to 64 normal subjects, could distinguish two modes of approach, one in which the perceptual, the other in which the conceptual, level predominated. Analogous response patterns were manifest on the Rorschach She could, further, statistically differentiate two groups of subjects from each other, those whose efficiency suffered from conflicting tendencies between the two modes, and those who successfully combined the two. Here we may have a lead to deeper-lying individual traits, to personality patterns in which the cognitive activity of the person is embedded. Thus, B. Candee[33] was able to make predictions regarding the mechanical or artistic aptitude of testees, by determining at which point in the Kohs block series the subject would change from a concrete pattern matching to a piece-meal construction response.

5. Before expanding further, it may be opportune to make the proposed principles more obvious by illustration. As a typical case in point, let me compare certain results of applying Henry Head's Ear and Hand test to normals and to aphasics. According to Pearson, Alpers, and Weisenburg,[34] the test proved diagnostically unreliable for aphasics, since mistakes occurred almost as frequently among the normal control group. A *qualitative* analysis, however, yields interesting results. In experiments carried out independently by H. Gordon[35] and F. Quadfasel,[36] it was shown that normal adults and children have a

[31] Cf **Long, L., & L. Welch.** Influence of levels of abstractness on reasoning ability. J. Psychol. **13:** 41–59. 1942; **Riess, B. F.** Genetic changes in semantic conditioning. Paper read at meeting of E P. A. June, 1945.

[32] A study of personal patterns in an intellectual performance. Char and Pers **9:** 315–25. 1944

[33] Personal communication

[34] Aphasia. A study of normal control cases. Arch Neurol. and Ps **19:** 2 1928

[35] Hand and ear tests Brit J Psychol **13.** 1923

[36] Ein Beitrag zum Motorischen Verhalten Aphasischer. Monatsschr. Ps u Neurol. **80:** 151–88. 1931.

genetically rooted, primitive, visuo-motor tendency to use *first* the hand which faces that of the experimenter, when imitating his movements. This explains the errors committed in *all* groups on a functionally plausible basis; for example, younger children tend to offer the left hand in shaking hands, *i.e.*, the hand which faces directly the other person's hand. Pearson, *et al.*, themselves mention that the cheer leader in the Red and Blue song of the University of P., while facing the stands moved to *his* right, whereupon the spectators moved to *their* left (*i.e.*, they did *not* reverse sides). Orders were then given that, if the song required moving to the right, the cheer leader should move towards his left. Similarly, to avoid confusion, the gym teacher in school moves towards *his* left when the students who face him should move towards *their* right, *i.e.*, both move to the same side. The experiments also provide, however, proof that normal adults and children above 8½ years can assume another attitude and *learn* to shift so that they reverse what they see, when imitating the examiner, and correct mistakes spontaneously. Both soon 'transpose,' by various ways and means. They either grasp the principle, 'All is in reverse,' or they place themselves imaginarily into the position of the other person, or they acquire an automatically operating, 'crosswise,' motor set. In contrast, the aphasic cannot change his original approach and learn to shift. He either does not recognize mistakes; or he makes futile efforts to correct them and to shift; or he succeeds but perseverates; or he succeeds partially, *e g*, in crossing the midline with the correct hand, he arrives at the wrong place, etc. (Head's and Quadfasel's records show that practically no patient succeeded on tasks involving "left hand to right ear".) It is crucial corroboration of this picture that aphasics, normal adults and children alike, make *no* errors when both the examiner and the subject look into a mirror and the subject imitates the examiner's movements which he sees in the mirror. (There is, of course, no reversal when looking into the mirror.)

I mention this problem in such detail, because it presents in a nutshell the issues raised, and concretizes the principles of technique here advocated. We are dealing with a functionally simpler level, on which normals and patients tend to operate naïvely. Therefore, both make *no* errors under one condition, in the mirror, where this approach on the concrete level suffices. Both do make errors in everyday life and on the test, where this naïve approach does not suffice for the task The crucial *difference*, however, appears when only the normal accomplishes the shift to the abstract approach, which is requisite for the

solution and for the process of learning 'reversals.' *In addition, we are dealing with various avenues* through which the shift is enacted, and these may be in turn characteristic of individual differences. *A propos* of the motor or visual imagery which here comes into play as an instrumental 'avenue,' it seems rather deplorable that research interest in this area has only recently reawakened.[37] We are not helped by the statistical proof that no imagery "types" exist, since we are in need of knowing which role imagery *patterns* play in the task solutions of different personalities.

6. On this basis, the term 'level' acquires a holistically defined meaning. In the *child,* the hierarchic order of lower and higher levels of function is chiefly conditioned by a *developmental* order. In the *adult*, lower and higher order of function is not necessarily *identical* with the genetic order of earlier or later. 'Level' becomes now a 'potentiality system,' of greater or lesser complexity, through which the whole person can operate in a more or less differentiated form. In the adult, it is largely the *present* situation, as the whole outer and inner field structure, that determines the type and degree of functional organization which is activated in a performance. We therefore have to distinguish qualitatively between the various *situational conditions* which make for the activation of higher or lower order systems. The symptomatic value of a performance hinges not alone upon the system through which the person has realized it, but, at the same time, upon the *role* which that system plays in the functioning total individual, under the given conditions. There are *organizational* differences as to how and why the 'same' system comes into relief. For example, the prevalence of a characteristically "primitive" system has *different meaning,* depending upon the framework in which it occurs. It differs in relation to one culture as against another; in relation to a lack of capacity in the genetic sense of 'not yet available,' as against an impaired capacity in the sense of 'no longer available'; in relation to a momentary frustration,[38, 39] as against a continued motivational aberration.[40, 41] Only with such distinctions in mind can we adequately evaluate performances as indicators of personality structure, normal

[37] Cf **Golla F., & W. Hutton.** The objective study of mental imagery. J. Mentl. Sci· 216–223. 1943.
[38] **Barker, R., T. Dembo, & K. Lewin.** Frustration and regression, etc. U of Iowa Stud. **18:** 1. 1941. **Dembo, T.** Der Aerger als Dynamisches Problem. Psych Forsch. **15:** 1–144. 1931.
[39] **Rosenzweig, S.** An outline of frustration theory. In **Hunt, J. McV.** Personality and the Behavior Disorders· 379–388. Ronald Press, New York. 1944.
[40] **Rickers-Ovsiankina, M.** Studies on the personality structure of schizophrenic individuals. J. Gen. Psychol. **16** (I and II): 153–178, 179–196. 1937.
[41] **Cameron, Norman.** Reasoning, regression, and communication in schizophrenics. Psychol. Monog. 1938. **50** (1): 1–34.

and abnormal, and can we, for instance, avoid the pitfalls of oversimplified "regression" or "repression" hypotheses. A statement of Babcock's seems germane, if we generalize it for our purpose: "Knowledge about separate mental functions is practically meaningless unless the interrelation with other factors and the relation to the functioning whole is known."[42]

This consideration entails the recognition that functional testing has to attempt even more than to provide crucial experiments and tasks, the solution of which requires processes which the experimenter can control. We also have to devise situations in which the method of approach and type of solution can only be actualized when they are necessary 'parts' of a definite motivational context. It is this attitudinal 'belongingness,' or, rather, motivational 'requiredness' of a given response, that the ideal test situation has to insure, a principle that could also be applied to the design of questionnaires. Therefore, not all tests used for the study of personality patterns are equally suitable for determining the different motivational backgrounds of responses. The content of questions or tasks has to be *pertinent* to the motivational processes for which we are testing. Beginnings in this direction have, for example, been made by Rosenzweig.[43]

7. I have previously alluded to the factor of learning in test situations. Goldstein and Scheerer[44] have recommended a method to test learning ability in the Kohs block performance and in the Weigl color form test. In the Kohs designs, concrete visual aids are introduced to teach the subject the procedure in case he failed. The criterion for final success is whether the subject can dispose of these crutches and has learned from these aids, when he is again confronted with the original task. I recommend that we should make wider use of the method of graded helps and graded 'pressure' (as in Klopfer's 'testing the limits') in mental testing. In other words, we should convert static testing into dynamic testing. We could make the test situation a learning situation, by measuring the extent to which the subject transposes what he learned to *other* situations, and so reveals his potential span of transfer. In this connection, Vigotski[45] has made the pertinent comment, "The ability to utilize the help given for the solution of a specific task is an important indicator of the subject's level of development, but is seldom taken into consideration by

[42] Cf. [11] l. c. 266.
[43] A test for types of reaction to frustration Am. J. Orthopsychiat. 4: 395–403 1935. See also the author's comments in the discussion of this paper, p. 678.
[44] Cf. [11] l. c.
[45] Quoted from **Hanfman, E., & A. Kasanin.** A method for the study of concept formation. J. Psychol. 3: 529. 1937.

the usual testing methods. Not utilizing this means of differentiation, however, may lead to a failure to detect the actually existing difference between two subjects or two groups of subjects."

Perhaps Vigotski underestimated the state of development in this field. I should like to remind you here of the work of the Spearman[46] school, especially of the carefully prepared studies by S. A. Laycok[47] and J. J. Strassheim,[48] in which "adaptability to new situations" was made the gauge of intelligence. British school-children were, for instance, presented with an imaginary situation and given a way of solving it. They were then placed in a number of similar situations, composed of increasingly different elements, to see how far they were able to apply what they had learned. In these new situations, the solutions could be reached only by varying the principle involved in the first solution. Interesting differences were found with regard to younger and older, duller and brighter, subjects. In the subsequent task variation, the duller children tended to reproduce in a rather mechanical fashion the responses learned in the first situation. Here "the eduction is one in which the relations are still in intimate contact with the fundaments."[48] The implications of these studies for the theory of intelligence and learning are outside of the scope of this paper.[49]

Perceptual-Motor Aspect

I should like to make a few comments on the possible use of experiments in this field, as indicators of personality. It seems to me, we have somewhat neglected to explore the problem of individual differences in perception, in favor of gross averages. We have grown too accustomed to accept perceptual laws on the basis of statistical majority, without showing scientific curiosity about the non-conforming minority. From the point of view of theory, however, we should feel obliged to account for both the majority *and* minority by an explanatory principle from which we understand the phenomena on both ends of the scale.

R. S. Woodworth,[50] *e.g.*, has taken exception to the mostly negative findings in experiments on reading facial expressions, because judgments were usually classified simply as right or wrong, without scrutiny

[46] **Spearman, C.** Creative Mind. Appleton-Century Co., Inc. New York. 1931.
[47] Adaptability to New Situations. Warwick Publishing Co. New York. 1939
[48] A New Method of Mental Testing. Warwick Publishing Co. New York 1926
[49] For related experiments and general discussion, cf. **Katona, G.** Organizing and Memorizing. Columbia University Press. New York. 1940.
[50] Experimental Psychology: 249. Henry Holt and Co. New York. 1938.

as to *how far* wrong they were. R. Arnheim,[51] when introducing the technique of matching (*e.g.*, handwritings, portraits) to the names of artists, simultaneously introduced "the analysis of errors" of the processes leading to these and a corresponding experimental check-up. This afforded a psychological explanation of both the correct and incorrect matchings, in terms of the structural characteristics of the matched contents and the personal difference in the attitudes of the subjects. Similarly, S. Asch has criticized the early experiments by Moore on suggestion, because Moore failed to investigate and to explain the behavior of those subjects who did not conform to majority opinion.

Especially in perception does it remain a tantalizing possibility that performance differences may provide clues for individual differences. We could explore the particular psychological and organismic systems which lie behind these differences. Among the many studies which have been undertaken in this direction, only a few remained unchallenged. The problem became all the more complex, as well as obscured, when the theories on constitutional types thrived in the era of Kretschmer and Jaensch. Both linked perceptual and motor activity of individuals to their constitutional type. When these notions of constitutional types exploded under the impact of contrary experimental evidence, the alleged findings on individual perceptual motor differences vanished with them.

I have often wondered why we have banished from our thoughts certain experimental results which Jaensch, Kretschmer, and their disciples claim. They could, at least, be given the benefit of the doubt, where they relate to perceptual or motor differences among individuals. I am thinking here of the following examples: B. Schmidt[52] reported that he could determine *preference for form or color*, by producing apparent motion of colored geometric figures where the factors favoring *phi*-movement in the direction of like color or like form were experimentally equated. F. Kranz[53] reported that impressive colors in space, with no cues for depth provided, are seen measurably closer by one group of subjects than by another. Further, in viewing for a short duration a vertical rod through prismatic spectacles which deflect the rod 15°, the

[51] Experimentalpsychologische Untersuchungen zum Ausdrucksproblem. Psychol. Forsch. **11**: 1–132. 1928.

[52] Reflektorische Reaktionen auf Form und Farbe. Ztschr. Psychol. **137**: 245–310. 1936; similarly, **Oeser, O.** Some experiments on the abstraction of forms and color. Brit. J. Psychol. **22**: I & II. 1932. See also **Thurstone, L. L.** A Factorial Study of Perception. Univ. of Chicago Press. 1944.

[53] Experimentell-strukturpsychologische Untersuchungen ueber die Abhaengigkeit der Wahrnehmungswelt vom Persoenlichkeitstypus. Ztschr. f. Psychol. Ergaenz. Bd. 16. 1930.

first group sees the rod *vertical,* the second does not. Similar differences were found with the Aubert phenomenon. Here, a vertical line is viewed in the dark room with the head in a laterally tilted position. The line then appears either tilted to right or left, in varying degrees to different individuals. Such experiments, which seem to differentiate between a more critical and more naive participating perceptual attitude among individuals, deserve repetition on a better controlled basis.

This recommendation receives emphasis through recently found personality differences among individuals who vary with respect to their adjustment to the change in the frame of reference of spatial orientation. After M. Wertheimer's[54] discovery that, when looking into a tilted mirror, not all subjects interpret alike the vertical and horizontal lines of the mirrored room, H. Kleint[55] made a series of brilliant experiments, in which he particularly studied the interrelation between postural tone and visual localization of directions in space. Even in these studies, we have unexplained 'minority' records, and neglect to follow up cues indicative of individual differences. The same holds for many otherwise fruitful studies on intersensory relations for which space does not permit discussion. There is, however, increasing realization that the problem of interdependence between sensory and motor functions contains one of the keys for an understanding of personality differences. This idea extends into new attempts by M. Rickers Ovsiankina[56] and H. Werner[57] to marshal experimental data for developing a theory of the psychological laws which underlie the Rorschach responses.

A promising lead for the investigation of personality through perceptual techniques is provided in a recent study conducted by Wertheimer, Asch, and H. Witkin, and later by Witkin.[58] These studies have been concerned with the question of how the individual gets and maintains his bearings in space under changing conditions. To investigate this problem, various techniques were employed: direction of the vertical-horizontal axes of visual space were changed by turning the visual field; the gravitational vertical acting upon the body was altered through the application of centrifugal force; the visual field was en-

[54] Experimentelle Studien ueber das Sehen von Bewegung. Ztschr. f. Psychol. **61:** 161–265. 1912.
[55] Versuche ueber die Wahrnehmung. Ztschr. f. Psychol. **138.** 1936; **140, 141.** 1937; **142, 143.** 1938.
[56] Some theoretical considerations regarding the Rorschach method. Presidential address, Rorschach Exchange. **7:** 2. 1943.
[57] Motion and motion perception; a study in vicarious functioning. J. Psychol **19:** 317–327. 1945.
[58] To be published. Cf. also **Gibson, J. J., & O. H. Mowrer.** Determinants of the perceived vertical and horizontal. Psychol. Rev. **45:** 300–323. 1938.

tirely eliminated by using dark-room situations and presenting luminous lines in different positions which had to be judged; and so on.

In the more than a thousand cases studied thus far, the most striking individual differences have been revealed. And the differences seem to be ranged along a single main dimension: the extent to which the individual relies upon the prevailing *visual* field or upon his *own body,* in establishing a frame of reference for orientation. In other words, the obtained differences are in terms of the individual's dependence upon, or independence of, the outer environment, in getting his bearings. So large were these individual differences that it did not seem possible to explain them on the basis of *specific* sensory features. Accordingly, it became necessary to consider broader aspects of the person to account for the origin of these differences, and, for this reason, personality studies were undertaken.

While this phase of the study is still in progress, the results already available indicate that the quality of orientation in the perceptual sphere seems to be correlated with the quality of orientation in the social sphere. Specifically, it is indicated by present results that people who depend very strongly upon the visual field, rather than upon their own bodies, in their space orientation tend to have personalities characterized by strong dependence upon others and low self-reliance. Thus, it is established that severe neurotics, in whom consciously experienced insecurity is the main trait, show a slavish dependence upon the visual field in their space orientation under the conditions of the test. It is significant, in this connection, that children are also very strongly dependent upon the prevailing visual field in their orientation, and women are significantly more dependent upon the outer field than men.

At the other extreme, where orientation performances involving a greater degree of independence of the visual field occur, we find a kind of personality in which well-founded self-reliance and independence seem the main traits. At the same time, there are indications that at this end of the distribution are also to be found personality patterns involving strong resistance to the outer world, such as is found in certain kinds of psychopaths. The dependence-independence dimension in the perceptual sphere is not correlated with health-unhealth, since healthy and unhealthy personalities seem to be represented at both ends of the distribution. These preliminary results, while undoubtedly oversimplifying the picture, strongly suggest that these perceptual performances do not represent narrow, isolated facets of the person's

make-up. It seems, rather, that through them are being tapped broader aspects of the person's characteristic relation to the world about him.

Perhaps these results bear a relationship to C. O. Weber's[59] interesting discovery, on 76 subjects, that extroverts make reliably larger corrections for size in a size constancy test than do introverts, *i.e.*, they show what R. H. Thouless[60] called a larger index of "phenomenal regression." Psychologically, this would again bespeak a more naïve 'object-directed' attitude and responsiveness to outer world influences. The latter experiments are also instructive from this point of view. It is often said that purely perceptual processes are too peripheral to signalize marked differences of individuality. It is, therefore, noteworthy that W. Köhler and H. Wallach,[61] in their research on visual after-effects, also emphasize the factor of great inter-individual differences How even so seemingly peripheral devices as tachistoscopic reading experiments can differentiate between normal and schizophrenic persons, is nicely illustrated in the study of A. E. Angyal.[62] She analyzed the omission and wrong letter substitutions, as well as the pattern consistency, in the reading sequence. On this basis, she arrived at statistically significant differences between both groups and between the paranoid and hebephrenic subgroups. All patients, taken together, make more errors, in terms of omission and unusual substitutions, than normals. However, the paranoids show almost exclusively omissions and a high pattern consistency ("pedantic rigidity"). The hebephrenics show high, unusual substitution ratios and a low pattern consistency (bizarre gestalt distortions, loose-shifting).

III. CATEGORIES OF BEHAVIOR

Modern students of personality have presented us with multiform attempts to overcome the atomizing classifications of behavior, and to do justice to the unique totality of the person. This is hardly the place to engage in any discourse on the multitude of personality dimensions as they have emerged from the organismic reorientation in psychology. Focusing upon actual and potential research, however, the

[59] The relation of personality trend to degrees of visual constancy correction for size and form. J. Appl. Psychol , **23** (6): 703–708. 1939

[60] Individual differences in phenomenal regression. Brit. J. Psychol. **22**: 216–241. 1932.

[61] Visual after effects, an investigation of visual processes. Proc. Am. Phil. Soc. **88** (2): 269–357. (cf. pp. 283, 346). 1944.

[62] Speed and pattern of perception in schizophrenic and normal persons. Char. and Pers. **10** (2): 108–127. 1942.

following selection of behavioral categories may be representative for our problem.

G. W. Allport has characterized behavior as adaptive, expressive, and projective.[63] This distinction has been experimentally elucidated in a recent study of L. Bellak.[64] He induced projection of aggression into certain T. A. T. pictures after he frustrated the subjects with severe criticism of their stories. Through control experiments and statistical analysis, he could ascertain that certain pictures elicited more often stories of aggression under *normal* circumstances than others. These same pictures induced higher rates of aggression responses in the *frustrating*, than in the non-frustrating, situation; also, more aggression than other pictures which did not suggest aggression by their actual content. Bellak concludes: "If the stimulus and the task are not well defined, a stimulus more suggestive of aggression will allow aggression to be projected more easily than one not suggestive of aggression at all." These results confirm previous work, for instance, by Levine, Chein, and Murphy.[65] Here, where subjects interpreted ambiguous pictures after prolonged hunger frustrations and when autism decreased, it became clear that projection is also a function of the real nature of the material with which a person deals, and is not simply arbitrarily imposed upon it. From this, it would appear that projection is at the same time adaptive behavior. In Bellak's case, the person copes with the task of interpreting, under duress, pictures of a vaguely structured content and imbues those which are objectively most suitable for this with aggressive characteristics.

We therefore make the preliminary generalization that behavior has the following aspects: 1. Intentful adaptive, in the sense of objective task orientation; 2. Unintentful expressive, in the sense of individual style and manner of task execution (or spontaneous activity); 3. Nonconscious projective, in the sense of *amalgamating ego* strivings with the contents of the task or of a creative product. Of course, projection can run the gamut from direct to indirect manifestations (*e.g.* from an attitudinal set or a patent wish fulfillment to compensating defense or to scapegoating). Thus, we may paraphrase Allport and Bellak by saying, adaptation and projection pertain to *what* one experiences or does, while expression pertains to *how* one experiences or *how* one does what one does.

---

[63] Cf. [4] and: The use of personal documents in psychological science. Social Science Research Council. 49: 111–124. 1942.

[64] The concept of projection. An experimental investigation and study of the concept. Psychiatry. 7 (4): 353–370. 1944. In presenting Bellak's findings, we do not accept his specifically Freudian definition of projection.

[65] The relationship of the intensity of a need to the amount of perceptual distortion. J. Psychol. 13: 283–293. 1942.

To take a concrete instance: an artist's work will be determined by the task goal he has set himself and the material with which he works: paint, clay, marble, tones, or language. Simultaneously, there will be his stylistic individuality expressed in the specific articulation of the formed material. Together with that, will be embodied a projection of *ego* tendencies, *ego* sensitizations which are reflected in such different psychological worlds and human situations as created by Balzac and Dickens, Proust[66] and Dostoyevski, Thomas Mann and Goethe.

On the experimental level, I may refer to Buhler's[67] application of the ball and field test to normal and abnormal children, and to the previously mentioned variations by Wells. According to Wells, the "systematic" approach which subjects evidenced in the more *detached, imaginary* search for the ball or purse was disrupted when the task was changed to a *concrete* one. It required searching for a concealed hole, in a paper covered board, with a pencil, and poking through the paper at the appropriate spot. "The presence of an actual goal with its accompanying frustrations introduces dynamic and affective factors that commonly bring a less rationalized response pattern." He speaks in this context of an adaptive compromise between responses to the task and to "subjective autistic factors," (cf., the impatience, anger, etc., manifested in looking for a lost object in real life). Bühler, analyzing the solution designs drawn by neurotic children, found "formalistic solutions of painful accuracy" together with "involved ornaments" and path confusion. This she did not attribute to a lack of understanding, but to obsessive trends, indecision, and evasiveness, since intelligence factors could be ruled out. Milder formalistic and ornamental trends, however, also appeared in perfect solutions of normal children. This reminds us of the observation by M. Mead[68] that 50% of the Samoan children tended toward an esthetic design on this test, and only a few subordinated this tendency to an objective solution. Whereas, in this case, projection was culturally conditioned, D. S. Porteus found that delinquent children could not successfully subordinate themselves to the instruction to trace the maze with the pencil without trespassing over the printed boundaries of the path.

All these illustrations taken together are offered in support of the further generalization that adaptive, projective, and expressive behavior are always coexistent. Which of these aspects will be in the *foreground* of the performance will normally depend upon the degree of

---

[66] **Heider, F.** The description of the psychological environment in the work of Marcel Proust. Char. and Pers. 9 (4): 295–314. 1941.

[67] The ball and field test as a help in the diagnosis of emotional difficulties Char. and Pers. 6 (4): 257–73. 1938.

[68] Coming of Age in Samoa: 291. Wm. Morrow Co., New York. 1939.

structuredness as to the meaning and realness of the *situation.* To this correspond the degrees of freedom it offers for the person. We may, for example, only compare the solving of an arithmetic problem or the driving of a car with dancing or with day-dreaming or with creative imagination. Projection or individual self-expression in automobile driving may be fatal. We are dealing, so to speak, with two poles of a dimension on which personality articulates. The less structured a situation, the more individual expression and projection will be allowed for, as far as the stimulus content is appropriate to either. The more structured the situation, the less expression and projection will be allowed for, and the more task-oriented, that is, the more adaptive, behavior will be prominent. The normally organized performance will, therefore, present a balanced patterning of the three aspects *fitted* to the reality valence of the situation. Where this adequate interrelation is disturbed, anyone of these aspects may be thrust disproportionally into the foreground and thereby alter the total pattern. Then we observe the neurotic deviations or the psychotic distortions in the disengagements from reality.

If this be true, then one may raise the question to what degree we are methodologically justified in *isolating* a particular projective or expressive aspect from the *total* behavior, by subjecting it to a specifically circumscribed test: the danger of creating artifacts because of the deliberate exclusion from the test situation of the normally coordinated other aspects should not be underrated. The arguments of William Stern against any "'monosymptomatic" diagnosis of personality,[69] and the methods explored by Allport and Vernon,[70] by Wertheimer, Arnheim, and W. Wolff,[71] should serve as a guide in this connection. The tendency of many psychoanalysts, in their verbal formulations at least, to treat unconscious motives as more real than the conscious motives by which they are supposedly concealed should serve as a warning. Similarly, even the "knowledge" of what is being projected, separated from the broader context in which the projection occurs, can be extremely misleading.

There are other reasons for being on guard against monosymptomatic isolation of one of these aspects. Our real concern is the organizational problem. How are these aspects related to, and centered in, the person as a whole, and what are the directional dynamics? In my opinion, it is insufficient to add another aspect and ascribe to it a

[69] Cf. p. 654.
[70] Studies in Expressive Movement. The Macmillan Company New York 1933
[71] The Expression of Personality. Harper and Brothers. New York. 1943.

vector quality, as would be the case if I now mentioned motivation. It seems to me that, in the final analysis, adaptive, expressive, and projective behavior can only be dealt with as 'parts' of the total adjustive process in the person's self-realization and coming-to-terms with his world. And this process is *co-extensive with evaluative behavior.* Just as much as the behavioral environment is co-extensive with meaningful relations and values, so is the self conjoined to these in actualizing its potentialities. Here, in evaluative behavior, we have the integrating personality dimension to which the various performance aspects are only subordinate, and of which they may even be "canalizations."[72] In proposing to replace the hypostatized concept of motivation by the category of evaluative behavior, I am guided by the conviction that *ego* values and objective values are not normally as discrepant as customarily assumed.[73]

Association psychology and psychoanalysis begin with an *ego* in 'splendid isolation.' Upon separation of the subjective drive contents as "*ego* values" from the objective values, it became a problem how the two are brought together; how to explain, for example, genuinely unselfish motives. This led to the constructs of *ego* expansion, of identification, *superego* and the like. From the position here advanced, the *ego* moves along a dimension where objective task orientation and *ego* orientation can intrinsically coincide, or can fall apart. In one case, we have an adequate balance between subjective and objective requirements. In another case, subjective and objective requirements either conflict or diverge. Thus, in the experiments of H. Lewis,[74] the question is not of *ego*- or non-*ego*-involvement, but of the *kind* of *ego* involvement. When task completion tensions are resolved in Lewis' subjects by someone else, it is not because the tasks are not meaningful to *them*, but rather because their *egos* are not involved in such a way that *they* have to complete the tasks, *e.g.*, for external reward or success. Both the subjective and objective requirements are satisfied when the task is completed, regardless of who completes it. We have, further, the comparable study of A. J. Marrow,[75] in which, despite overt completion, the subjects experience the task as incomplete, and, despite overt interruption, the subjects experience completion, with resultant differences in recall. The picture can be restated in simpler language.

[72] Cf. for this term, **Murphy, G., L. B. Murphy, & T. M. Newcomb.** Experimental Social Psychology. Harper and Brothers, New York. 1931.

[73] Cf. **Köhler, W.** The Place of Value in the World of Facts. Liveright Publishing Corp. New York. 1938.

[74] An experimental study of the role of the ego in work. J. Exp. Psychol. **34** (3): 113–126; 195–215.

[75] Goal tensions and recall. J. Gen. Psychol. **19** (I and II). 3–64. 1938.

In having the experience of finished or unfinished 'business,' our *ego* is more task-involved than self-centered.

On the other hand, the *ego* may be involved in so self-centered a way that harmony with the external situation becomes impossible. This is the essential dynamics of frustration tolerance. Tolerance is low, if the understanding of the situation is such that the threat to the individual intensifies self-centeredness and detracts from task involvement. This, for example, is the case when external pressures of success or failure, of competition and status, make the *ego* "reactive" to features of the situation which are extrinsic to the nature of the task, or when the task demands overtax the abilities of the person. Frustration tolerance is high, if the understanding of the situation is such that the threat to the individual is meaningfully connected with the realization of the task and his potentialities; or that the threat has no bearing upon the *ego*-task-centered orientation.[76] Thus, the soldier in battle carries out his task by tolerating deprivation and the threat of death or mutilation; the public speaker, the actor overcome "stage-fright." The experimental studies of Rosenzweig[77] and Lewis[78] illustrate the point that ego-task-orientation with over-emphasis on the *ego* affects recall negatively

In this light, Allport's[79] introduction of the *ego* to contemporary psychology seems only partially adequate. He insists that *ego*-involvement (even with external praise or reproof) is of importance to oneself in practically every sense, and that *ego* satisfaction is a decisive factor in motivation. Thereby, he seems to overlook the important balance between *ego*- and task-orientation, and tends to develop a conception, not so much of *ego*-involvement, as of *ego*-over-involvement. He even seems to go so far as to regard such involvement as prerequisite for adequate performance.[80] The problem of performance adequacy is not simply one of *ego*- or non-*ego*-involvement, but rather of the degree to which the *ego* can *effectively* participate in a task. In Allport's most recent paper,[81] however, the balance seems to have been improved. Here we find the statement: "Perhaps the most important distinctions concern reactive *ego*-functions, which are resistant,

[76] Cf. **Maslow, A. H.** Deprivation, threat and frustration Psych. Rev. **48** (4) 364–366. 1941.

[77] An experimental study of repression with special reference to need-persistive and ego-defensive reactions to frustration J Exp. Psych. **32** (1)· 64–74. 1943.

[78] An experimental study of ego-orientation in work. Paper delivered at EPA meeting. 1945.

[79] The ego in contemporary psychology Psych. Rev., **50** (5): 451–478. 1943.

[80] Note, however, **Allport's** remark (p. 467) that "Too intense an ego involvement may be disruptive" The paragraph containing this remark, however, seems to be almost a parenthetical reservation when set against the rest of the material which he introduces.

[81] The psychology of participation. Psych. Rev. May, 1945.

contrary, clamorous, as opposed to active *ego*-functions which find full expression in participant activity. When participating, the individual discovers that his occupational manipulations grow meaningful . . . in studying participation, the psychologist has an approach to the complete person." With this change in emphasis, the *ego* in contemporary psychology has emerged from solitary confinement and become a participating *ego*. What remains is to recast "*ego*-involvement" and "task-involvement" in terms of evaluative behavior; that is, the adjustment of the person to the total situation, which encompasses subjective and objective requirements. For every adjustment establishes a value, and it is only in terms of evaluative behavior that we can appreciate adequate or inadequate subjective or objective emphases.

What are the implications for the study of performance? Space does not permit us to pursue this question. Its ramifications lead to the problem of how *ego*-organization and the structure of the social field are related. One conclusion, however, immediately presents itself. We have to investigate the *field* conditions which are conducive to adequate integration of *ego*-involvement and task-orientation, and those which are not. This is not a question "of the optimum degree of *ego*-involvement," but of *ego*-task-integration. From here, we can develop the important practical applications for fostering normal, and preventing abnormal, self-actualization.

DISCUSSION OF THE PAPER

Joel Shor (A. U. S.): The papers of Dr. Rapaport and Dr. Scheerer offer much theoretical clarification of our clinical procedure in diagnostic testing of military neuropsychiatric casualties We give each man at least four tests, including a Wechsler scale, a verbal projection sheet (*sic*), drawings, and often a Rorschach or Thematic The combined use of the projective and non-projective features on each instrument provides data which we organize into statements about both the basic drives and impulses, and the quasi-stabilized or conventional response patterns in meeting reality challenges ("the modes of controlling the basic impulses," as Dr Rapaport put it, or "coping mechanisms," in Dr. Scheerer's terms) While these papers have spoken mostly about schizophrenics and aphasics, we find the general run of psychoneurotics may also be effectively examined in these terms. The bogey of "the inspired guesser" is still with us and only inspires to refine and further validate our procedures.

At least, one will see clearly from these papers that clinical psychology has advanced, out of the stage of the I. Q. and artifactually isolated personality traits, towards an integration of data on the development of personality and the psychology of thinking And Dr. Rapaport's formulations have not omitted the cultural process of conventionalization, the sociological framework. However, we have yet to apply analytic findings concerning the dramatic course of ego development and the correlating development of ego functions. Growth curves for specific intellectual processes may no longer be assumed to match the "total mental development curve" so flagrantly advertised in texts of the last two decades. As for personality traits, we may still use tests of self-evaluation, interests, and prefer-

ence statements. However, these data are then made meaningful in the light of an investigated motivational framework and of a more basic understanding of personality forces and balances; not as a valid report *per se*.

Dr. Morris Krugman (*Bureau of Child Guidance, Board of Education, New York*): Since clinical psychology is suffering from too much empiricism, and too little sound theory, both papers are very timely and appropriate, as the culmination of the numerous research and practical reports presented here.

The term, "non-projective techniques," as used throughout these papers, is rather puzzling. It seems to be a reservoir term for all psychological instruments except the projective techniques. Since a bibliography of psychological tests is likely to contain more than five thousand titles, and since only a handful are projective techniques, the non-projective techniques constitute all but a fraction of one per cent of psychological instruments. This is very much like classifying people as feebleminded and non-feebleminded, or psychotic and non-psychotic. The term used, however, is unimportant, but the concept is important. Dr. Rapaport distinguishes between structured and unstructured material for studying personality. This differentiation between the non-projective and the projective techniques does not give the full difference between the two; Dr. Rapaport's further criterion, that in the projective technique the subject does not know what is expected of him, is even more important for a non-projective technique. If this criterion is met, either in the non-projective techniques or in the projective, the difference between the two is almost completely eliminated. Thus, in a test battery like the Wechsler-Bellevue Adult Scale, the material is structured, and the test measures intelligence. However, if we pattern the responses for personality diagnosis, or use differences between non-stationary functions and quasi-stationary functions, as Dr. Rapaport suggests, then we are using the test so that "the subject does not know what is expected of him." The same test thus serves two distinct purposes, and, in effect, constitutes two different instruments.

The use of tests in the manner suggested by Drs. Rapaport and Scheerer is not new, although they have made important contributions. For many years, psychologists who were more than psychometricians have utilized clues from psychological instruments for an appraisal of personal and emotional adjustment. In the Stanford-Binet, for example, they used the methods of observation under test conditions, qualitative evaluation of responses, analysis of test variability, and clinical judgment *versus* objective results, evaluation of behavior during different parts of the test, analysis of verbal content, test patterning, intuitive evaluation, etc. Under these, specific items, like motor coordination, pathological slowness, impulsiveness, bizarre content, and many others, were observed. In effect, these approaches constituted the application of the projective technique principle; that is, the real purpose of the test was disguised to some extent. The importance of the work described at these meetings lies in the attempts at validation of well-known and fairly old techniques, and the removal of intuitive methods. This is not without some danger, however. Tables and formulae, thus developed, will be used mechanically and blindly by some for arriving at important clinical decisions. It becomes necessary to emphasize that, while these approaches are important adjuncts to clinical methods, they do not take the place of sound clinical judgment. Mechanical application of diagnostic tools should be discouraged.

The hypotheses or principles presented by Dr. Rapaport are extremely important; if we abide by them, they will clarify our thinking and make clinical diagnosis more valid. One of the hypotheses may require further elaboration. Dr. Rapaport stated: "Deviation from the trend of the general population is the individually, and the pathologically characteristic indicator." This is only part of the story. We must not forget the existence of individual differences: both inter-individual and intra-individual. A low performance score, for example, and a high verbal score may at times be diagnostically significant for schizophrenia, but there are times when such discrepancies mean simply that the individual in question has poorer manipulative, than verbal, ability. This is the type of error

some workers in projective techniques make when they use these instruments for diagnoses without support from other clinical sources. Workers with "non-projective" techniques should attempt to avoid this type of error.

One further consideration: The emphasis in most of the studies reported in these papers has been on the negative aspects of personality and on pathological factors. In wartime screening, this is necessary, but for clinical aspects in general, more emphasis is needed on the positive aspects of personality, and on the "normal." There is great need for research along these lines.

There is still need for more research with present techniques. We have hardly begun to extract the possibilities of existing instruments. The approaches reported at these meetings show the way admirably. They objectify methods of diagnosis which were formerly intuitive and largely subjective. In addition to research on new techniques, whether projective or non-projective, intensive work with present instruments should yield dividends in clinical diagnosis.

The questionnaire methods reported in the earlier papers seem to work well in the armed services, because conditions there are conducive to truth telling, and this is the basis of the questionnaires or the personality inventories. In the selective service process, or in the armed forces, the subject who wishes to be excused from military service, or to be discharged from the army or navy, has every reason to answer the neurotic inventory truthfully, since such replies would result in action desired by the subject. I have some doubts that the direct approaches to the study of personality would work effectively in civil life, since, in most instances, conditions are not conducive to truth telling. In fact, quite the reverse is true in most instances. (Such failure to tell the truth may be conscious or unconscious.) A candidate for a position, or for promotion, can hardly be strongly motivated to answer a personality inventory truthfully, if so doing means failure to obtain the position or the promotion. It would seem, therefore, that indirect approaches, whether projective or non-projective, hold most promise for personality study in civil life.

Dr. Scheerer's plea for an organismic approach to the study of personality can only be applauded. His statement that emoting, thinking, and conceiving are coexistent, and that a "performance" is a manifestation of the organism as a whole, and not of one part of the organism, constitutes a dynamic clinical point of view. Dr. Scheerer makes a plea for the analysis of process in psychological instruments. This is theoretically sound, but presents tremendous difficulties in test construction Most test items possess much overlap, and, at present, it is an almost impossible task to construct test items that segregate single psychological functions

There is one further question in this connection: If a thorough-going analysis of process is made in the case of each test item, does this not lead to a departure from the organismic approach? In order to avoid a reversion of the atomistic approach to the study of personality, it seems necessary to explain to what extent such an analysis of function is only for purposes of understanding and discussion, and not for personality evaluation. Unless this is done, the conclusions with reference to item analysis and organismic approach would seem contradictory

Dr. Rapaport: I shall restrict my discussion of the apparent contradiction in our referring to the "reaction of the organism as a whole," and yet speaking of testing distinct "functions" by different intelligence and concept-formation tests. I submit that, even though we speak of the reaction of the organism as a whole, we are entitled to treat of "discrete functions," as long as we recognize that what we call functions are merely our different ways of looking at psychological functioning, and that treating of psychological functioning is one of our ways of looking at organismic functioning. Thus, "memory," "concept formation," etc., should be considered different *aspects* of—different ways of our looking at—psychological functioning. Treating memory function *per se* is warranted when the phase of psychological functioning investigated is such that treating it from this point of view is the most expedient. It appears that certain phases of psychological functioning can be treated expediently from a certain point of view only, and that others have to be treated from several points of view at once, though there are none which can be exhaustively treated from only one point of view. When we say that the

function underlying "Similarities" on the Bellevue Scale is concept-formation, by that we merely state that it is outstandingly expedient to treat the phase of psychological functioning involved in responding to this test, from the point of view of concept formation. I do not see any contradiction in the organismic point of view and discrete functions being treated of in such manner.

Dr. Scheerer: I am asked whether and how the organismic approach permits the application of quantitative methods to personality study. My answer to this is that quantification of data is only meaningful if we first know what the data signify within their context. Therefore, qualitative analysis has to precede, or at least be conjoined with, quantitative evaluation. I think some of the studies I have mentioned in this paper bear testimony to the possibility and the fruitfulness of such approach.

On the other hand, this type of approach is naturally more difficult. For this reason, I have emphasized the study of cognitive aspects of performance and particularly experiments in the field of perceptual motor functions, because, in both, quantification of qualitatively analysed data seems more readily attainable. As I tried to show, some work has already been done successfully and the beginnings are promising. For the same reasons, I thought it necessary to suggest further studies in these fields, because it appeared to me that we have grown to minimize individual differences in perceptual motor functions and to overemphasize projective techniques.

Regarding the questionnaire methods in personality study, I am inclined to be more skeptical. Here we are usually in the situation where we cannot unequivocally evaluate the meaning of a given response as to the motivational context to which it belongs, because we have no adequate provisions to analyze the response qualitatively. In other words, we need experimental designs to focus the psychological processes underlying the questionnaire responses and to determine the whole motivational system in which these processes and the responses function.

Certain questionnaire methods to which I referred seem to represent an improvement in this direction; the nature of the questions, to begin with, defines more clearly situations as, for example, in Lickert's Technique or in the Rosenzweig frustration test. In this connection, other studies come to my mind which illustrate the importance of qualitative analysis, also, for questionnaire method. I am thinking of a paper by Paul Lazarsfeld on "The Art of Asking Why"; of the studies of Lewis and Asch where the respondents had *to give reasons* for the "how" and "why" of their answers; and of a study by Eisenberg. It appears to me that we can only succeed with questionnaire methods if we include the question, "Why?", in order to discover the particular context in which the answer functions for the given person. Eisenberg, for example, applied a neurotic inventory to a group of subjects, and then scored the subjects for absence or presence of neurotic traits. In a follow-up study, he included the question, "Why?," for the given answers. As a result, certain subjects who came out "neurotic" on the inventory had to be re-scored as normal, and *vice versa*.

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES

VOLUME XLVI, ART. 8. PAGES 679–882

SEPTEMBER 16, 1946

LYMPH*

BY

PHILIP D. MCMASTER, ROBERT CHAMBERS, ELIOT R. CLARK, THOMAS F. DOUGHERTY, CECIL K. DRINKER, WILLIAM E. EHRICH, EUGENE M. LANDIS, VALY MENKIN, PAUL A. NICOLL, RICHARD L. WEBB, ABRAHAM WHITE, AND B. W. ZWEIFACH

CONTENTS

* This series of papers is the result of a Conference on Lymph held by the Section of Biology of The New York Academy of Sciences, April 13 and 14, 1945.

Publication made possible through a grant from the Conference Publications Revolving Fund.

INTRODUCTION

By Philip D. McMaster

The Rockefeller Institute for Medical Research, New York, N. Y.

From time to time, conferences are held by the New York Academy of Sciences to afford an opportunity for the clarification of current knowledge of certain subjects. At each conference, a few men, actively engaged in research upon some common theme, present their work before a selected group with similar or allied interests. By free discussion, the combined knowledge of all is employed to arrive at better understanding.

The papers and summarized discussions which follow represent the matter presented at a conference on Lymph. The sequence of the papers and the subjects touched upon deserve a word of comment, since it is to be noted that as much has been written of blood capillaries, of interstitial fluid, of antigens and of antibodies, as of lymph.

A conference on lymph might be expected, by some, to include all of the anatomy, physiology, and pathology of the lymphatic system, as well as the chemistry of the fluid itself. Such a program would require many conferences. In our present state of ignorance, it seemed unwise to attempt such an all-inclusive program and unwise to try to present even a compendium of knowledge of what lymph is like in one state or another or of how it changes under various known conditions.

It seemed wise to be more humble and to attempt a consideration of only a few of the fundamental aspects of the subject, in order to form a common ground from which we might take departure for future work. Consequently, whole sections of lymphatic physiology and pathology are not represented in the titles on the program. For example, ever since the first discovery of the lymphatics, much of interest has been written about the part they play in the transport of fat and in the spread of new growths. A consideration of these matters would have taken this conference too far afield and would have left us in possession of a group of isolated facts in need of correlation.

Instead, it has seemed fitting, for the present conference, to ask Dr. Chambers, Dr. Zweifach, and Dr. Landis to discuss the anatomy and physiology of the blood capillary bed, in order to present, upon as broad a basis as possible, the origin of the fluid from which lymph is eventu-

ally formed. Further observations of the minute blood and lymphatic vessels in the bat's wing have been described by Dr. Nicoll and Dr. Webb. Dr. Clark has discussed the state of affairs in the interstitial tissues Some of the factors and forces affecting the transport of fluid through the tissues, in its conversion to lymph, and certain physiological and pathological conditions affecting changes in lymph flow have been described in the author's paper, as well as the relation of the lymphatic system to the spread of infection through its channels and to the defense of the body by the production of antibodies within regional lymph nodes. Dr. Menkin has discussed the changes taking place in lymphatics within inflamed tissues, and Dr. Drinker has reviewed his findings, which emphasize the function of the lymphatics in the take-up of protein from the tissues. Finally, recent observations on the role of the lymphocyte in the production of antibodies and in the processes of immunity have been described by Dr. Ehrich, and by Dr. White and Dr. Dougherty.

FUNCTIONAL ACTIVITY OF THE BLOOD CAPILLARY BED, WITH SPECIAL REFERENCE TO VISCERAL TISSUE

By Robert Chambers and B. W. Zweifach*

Department of Biology, Washington Square College of Arts and Science, New York University, New York, N. Y.

The physiological significance of the capillary circulation has engaged special attention, since the noteworthy contributions to the subject by Dale and Richards[1] and by Krogh.[2] Unfortunately, little consideration has been given to the possibility that the functioning of the capillary bed depends upon a highly specialized, structural organization of the vascular components. Except for the disputed Rouget cell concept, championed by Krogh, physiologists have tended to base their ideas on the existence of a network of capillaries of essentially uniform composition. in spite of the fact that the literature contains many allusions to real differences among the several components of the bed.

One of the difficulties in evaluating the published data is the indiscriminate use of the term, capillary. Bayliss[3] appreciated this. His criticism, that the capillary blood vessel still awaits definition, is almost as true today as when he made it, over twenty years ago. Another source of common confusion has been to apply the term in its literal sense (a hair-like structure), and to use caliber alone as a criterion. Actually, the components of the capillary bed can be classified under several categories of function. Moreover, the muscular components are often narrower than most of the true capillaries in the bed. There has been a tendency to make broad generalizations from observations on a few isolated vessels and even segments of vessels.

The use by Landis[4] of the micromanipulative method added much to our knowledge of the subject, but further investigations of this nature are necessary. The more recent micromanipulative studies of Zweifach[5, 6] have gone far in this direction and have demonstrated that the functional activity of the capillary bed depends upon a distinctive architectural pattern of morphologically and functionally different components.

* Present address: Department of Medicine, Cornell University Medical College, New York, N.Y.

An essential basis for understanding the blood capillary system is to recognize that the distinctive pattern of the bed varies in different regions of the body, according to the functions performed. The basic function is nutritive. However, in certain organs in which specialization predominates, the resulting deviations tend to mask the nutritive pattern. It is not possible, therefore, to present a generalized scheme. A description of any selected region must take into account the degree or lack of specialization to which the capillary bed is subjected. This is obviously true for such organs as the liver and the spleen. It is also true for the much-studied web of the frog's foot, of which the capillary circulation is markedly atypical, consisting mainly of haphazardly arranged, inter-anastomosing capillaries, with extremely pervious walls. A more strictly nutritional type of capillary bed is characteristic for the circulation of the visceral and muscular tissues Even here, a certain degree of variation exists.

The purpose of this report is to present the capillary bed as an organized unit, based on a description of the more generalized nutritional type. A detailed presentation of this has been given recently[7] and is summarized in TABLE 1.

Muscular elements have been found to be present in the capillary bed proper. However, instead of being indiscriminately distributed, as indicated by Krogh and his supporters, they are restricted to the well-defined, capillary-like, central channels and their precapillary off-shoots. The channels serve as thoroughfares from arteriole to venule. They are relatively long vessels and bear a close relationship to the true capillaries. Under normal conditions, the central channels remain open, so that any spontaneous restriction is caused by contraction of the precapillary off-shoots and by the recurrent vasomotion of the metarterioles. The vasomotion controls the rate of flow through the central channels, while the alternate opening and closing of the sphincters of the precapillary off-shoots induce an intermittent flow through the true capillaries, without interfering with the flow in the central channels. Even during the intervals when their supply of blood is shut off, the capillaries generally contain a fluid. This fluid, diffusing in from the interstitial spaces, is carried by way of the post-capillaries into the distal segments of the central channels and the venules.

A significant feature of the muscular components of the capillary bed, particularly the metarterioles and precapillaries, is their reactivity to epinephrine and to nerve stimuli, and their susceptibility to local

changes in the condition of the tissue in which they lie. When the tissue is in a state of comparative rest, the periodic phase of vasomotion is augmented. This results in a continuous flow only in the central channels and a slowed, sporadic blood flow in the true capillaries. Resting tissue is thus relatively ischemic. When the tissue is activated, as by mechanical irritation, so that conditions arising from the tissue predominate, the vasomotion ceases and the metarterioles and pre-capillary sphincters become dilated. Thereupon, the

TABLE 1

VASCULAR COMPONENTS OF THE TERMINAL CIRCULATION IN THE RAT MESOAPPENDIX

Dimensions given are of diameters.

I. **Terminal Arteriole** (ca. 20–25μ)
Coat: muscular, single, continuous layer.
Movements: pulsations related to larger arteries, but less regular.

II. **Capillary Bed**

1. *Metarteriole* (ca. 8–15μ) (initiating thoroughfare a-v channel).
Coat: muscle cells, typical, discontinuous.
Movements: Vasomotion, with slow, constrictor-dilator phases.
Branches: Precapillary Junctions (ca. 5–12μ) twisted, outflowing.
Coat: muscle cells, typical, acting as sphincters.
Movements: Vasomotion, independent of metarterioles.
Lead into *True Capillaries.*
Coat: endothelial.
Movements: passive and tonic.

2. *Proximal Segment of a-v Channel.*
Coat: muscle cells, atypical.
Movements: No vasomotion, responsive only to abnormal stimuli.
Branches: a. *Precapillary Junctions.*
Coat: muscle cells, atypical.
Movement: Similar to a-v.
Lead into *True Capillaries.*
b. *Capillary Junctions,* outflowing, same structure as and continuing as *True Capillaries.*

3. *Distal Segment of a-v Channel.*
Coat: connective tissue.
Movements: passive, slight.
Branches: Postcapillary Junctions, inflowing, same structure as and originating from *True Capillaries.*

4. *Non-muscular Venules* (ca. 15–25μ) (fusion of thoroughfare channels).
Coat: Pronounced connective tissue.
Movements: passive, slight.

III. **Muscular Venules** (ca. 25–30μ)
Coat: muscle cells and connective tissue.
Movements: Highly responsive, varicose contractions.

TABLE 2

TERMINOLOGIES IN LITERATURE OF COMPONENTS OF CAPILLARY BED

		Chambers and Zweifach[7]	Zweifach[5]	*Zimmerman[8] †Midsuno[9]	Clark[10] Hooker[11]	*Lewis[12] †Heimberger[13] Schneider[14]	Benninghof*[15] Ebbecke[16] Killian[17] Krogh[3]
Muscular	continuous	Arteriole	Small Artery	Arteriole	Arteriole	Arteriole	Arteriole
	discontinuous	Metarteriole	Arteriole Prox. a-v capillary	*Precapillary arteriole †Precapillaries	Arteriole	*End arteriole †Precapillaries	Arterial Capillaries
		(*Offshoots*) Precapillaries		Precapillaries	Arteriole	†Precapillaries	Arterial Capillaries
Non-muscular		True Capillaries		Arterial Capillaries		Capillaries	Arterial Capillaries
		True Capillaries		Arterial Capillaries		Capillaries	Capillaries
		True Capillaries		Venous Capillaries		Capillaries	Venous Capillaries
		Postcapillaries		Venous Capillaries		Capillaries	Venous Capillaries
		Distal a-v Channel	Distal a-v Capillary	Venous Capillaries		Capillaries	Venous Capillaries
		Non-muscular Venule		Postcapillaries	Venule	*Collecting Venules	*Postcapillary Venules
Muscular		Muscular Venule		Fine Venule	Venule	†Venule	Venule

Fulton and Lutz[18] refer to the precapillary as junctional capillary. Schaly[19] and Tannenberg[20] refer to muscular cells (*Pförtnerzellen*) about this vessel.

*† The superscripts relate the authors to the nomenclature used by them.

capillaries become flooded with blood, and hyperemia results. The metarterioles and precapillaries form an integral part of the capillary bed proper and their reactive muscular elements are so disposed as to be fully exposed to chemical changes in the environment.

It would appear that, at least in the visceral tissues, the functional autonomy of the peripheral vascular system depends largely on two features. One is the dual character of its components, *viz.*, a central or thoroughfare channel, of which the true capillaries are off-shoots. The other is the intermittent vasomotion of the muscular or metarteriolar portion of the central channel, and the muscular sphincters of the precapillary off-shoots which lead into the true capillaries. Both of these features have been mentioned in the literature, although fragmentarily and with certain variations in the nomenclature. TABLE 2, with references,[8-20] is a summary of the nomenclature used in the literature.

Previous to Krogh,[2] who makes no reference to the existence of a main channel, at least three investigators, Klemensciewicz,[21] Jacoby,[22] and Wolheim,[23] not only differentiated between main and accessory channels, but correctly interpreted their significance Klemensciewicz and Jacoby made their studies on various tissues of frog and described main channels, termed by Jacoby "*Stromkapillaren,*" which were always flowing, while, most of the time, the side-branches contained little or no blood. Klemensciewicz also observed that, during ischemia, the orifices leading into the side-branches are closed. Wolheim, in his observations on the human skin in which the blood vessels were made visible by blistering off the epidermis, distinguished between "*Stromkapillaren,*" in which the blood flow is rapid and uninterrupted, and "*Netzkapillaren,*" in which the blood is relatively stationary. The investigations of Zweifach, presented in a monograph[24] on various tissues of the frog and mouse, brought out most clearly the widespread occurrence of the main, arteriolar-venular channels, of which the true capillaries are side-branches.

There is also evidence in the literature concerning the periodic vasomotion of the metarterioles and of their precapillary side-branches. The existence of an intermittent flow in the capillaries has been repeatedly reported. Hagen[25] observed a shift in flow from one capillary to another, in the rabbit's ear. Krogh[2] referred to several investigators who observed intermittent flow, but attributed it to varying contractile responses of the capillaries. Relegation of the role of inducing capillary intermittency to the arterioles (our metarterioles) feeding the capillary

bed was made by Nesterow[26]; and by T. Lewis.[12] Wyman and Tum Suden[27] presented evidence for a recurrent vasomotion of the smaller arteries, from their observations on the serosal vessels of the rat's intestine, exposed through a window in the wall of the abdomen. They reported that the periodic changes in caliber occurred at intervals of about several seconds. They were uncertain whether or not the observed oscillations were a consequence of the operation exposing the tissue. Grant[28] had similarly reported periodic alterations of the small arteries in the rabbit's ear, which, however, were noted only after mechanical injury or after histamine injection. The existence of normal, rhythmic changes in caliber of the peripheral circulation as a whole has been indicated in plethysmographic records of the digits of the hand by Hertzman,[29] and by Burton and Taylor.[30] The investigators who have come closest to the type of vasomotion described in this paper are Sandison,[31] and Clark, Clark, and Williams[32] who obtained direct observations of arteriolar rhythmic contractions through transparent chambers in the rabbit's ear. The variable length of the periodic intervals reported by them (from a few seconds to 2–3 minutes) and the dependence of the contractions on nerve control have been fully confirmed by our observations on the mesoappendix of the rat.

Zweifach,[6] in observations of frog and mouse tissues, showed not only a relationship between the central a-v channel and its capillary side-branches, but the importance of the metarteriolar vasomotion and that of the precapillaries in controlling the flow of blood through the capillary bed.

It should be emphasized that the central a-v channels of the capillary bed are not to be confused, as was done by Wiggers,[33] with the arterio-venous anastomoses. A good review of the latter has been made by Clark and Clark,[34] and Clara.[35] The central channel of the capillary bed is a relatively long vessel, with capillary side-branches along its entire length. The A-V-A, on the other hand, is typically a relatively short, muscular vessel serving as a shunt, which crosses abruptly from an artery to its corresponding vein. The A-V-A are of several orders, according to the size of the artery and vein which they join. The A-V-A of the smallest order are those which bridge across from a terminal arteriole to its corresponding venule, where the latter leaves the capillary bed. These arteriolar-venular anastomoses are probably the structures referred to by Wright[36] as intra-capillary anastomoses. When wide open, the A-V-A may effectively cut off the circulation from the capillary bed. From the standpoint of the

general circulation, the widespread distribution of the A-V-A offers an effective mechanism in the terminal circulation to insure an adequate venous return and, thus, to counter-balance the effect of an abnormally increased peripheral resistance.

The tendency of the capillary bed to react independently of the rest of the vascular system has given rise to considerable study on the existence of contractile mechanisms in the capillary bed. The divergence of opinion in the literature, regarding this, lies chiefly in the lack of appreciation of the functional relationship between muscular, central channels and the capillaries into which they feed. Some investigators have contended that contractile muscle elements are distributed throughout the bed, while others have stressed the contractile nature of the endothelium. Krogh and Vimtrup,[37] Bensley and Vimtrup,[38] and Schaly,[19] among others, have claimed that the capillaries possess muscular elements (Rouget cells). Among the more recent to support the Rouget cell concept are Fields[39] and Beecher.[40] On the other hand, Clark and Clark and their co-workers (Clark and Clark,[10] Sandison,[31] Clark and Clark[41]), from observations of vessels in the transparent chambers in the rabbit's ear, over long periods, found no evidence of active contractility in any of the true capillaries.

Special mention should be made of the precapillary junctional muscle cells. They were called *"Pfortnerzellen"* by Tannenberg,[20] who observed them in the rabbit's mesentery; by Schaly,[19] in fixed preparations of the mammalian eye; and by Fulton and Lutz,[18] in the living retrolingual membrane of the frog. Fulton and Lutz described a periodic opening and closing of the junctions as occurring spontaneously, also in response to stimulation of nerves with microelectrodes. Similar precapillary sphincters have been noted in highly specialized regions. Richards and Schmidt[42] found them at the arteriolar source of the glomerular capillaries in the frog's kidney and remarked on the absence of contractility in the capillaries proper. Crawford[43] observed contracting mechanisms, at the arteriolar origin of the capillary loops, in the epidermal papillae at the base of the human fingernail.

Of interest, in this connection, are the observations of Sanders, Eberth, and Florey[44] on the reorganized tissues (46 days' growth) in the transparent chamber in the rabbit's ear. They claimed to have observed active contraction of the capillaries, with obliteration of their lumina by in-bulging of the endothelial cell-nuclei, with no decrease in out-

side diameter of the capillaries. Their published photographs show this to occur on abrupt offshoots at their junctional segments, indicating that the vessels are probably the muscular precapillaries.

Spontaneous movements of the endothelial cells were remarked upon by Midsuno.[9] He claims to have observed alterations in the lumina of capillaries in the frog by amoeboid movements of the endothelial cells. In our observations on the formation of endothelial ridges, during partial and rhythmic contractions of arterioles and precapillaries, we were at first struck by what seemed to be spontaneous changes in form and shape of the endothelial cells, producing elevations and folds projecting into the lumen of the vessel. However, the folds were always longitudinal and only appeared during the contraction of the surrounding muscle cells Kahn and Pollack[45] have described the longitudinal in-folding of the endothelium, as a result of the active contraction of pericapillary cells, induced by faradic stimulation on the excised nictitating membrane of the frog. However, they also described an inward swelling of endothelial nuclei, where they could find no evidence for pericapillary cell contraction We have observed similar conditions of the endothelial cells of true capillaries, but only during the absence of flow, a condition which can be ascribed to diminished intracapillary pressure, without assuming active contraction of the endothelial cell.

The endothelial cell has the ability for inherent, although restricted, movement, but this type of action has never been observed as being sufficient to occlude the lumen during blood flow. The cells sometimes exhibit amoeboid processes which rise and sink. A striking phenomenon is the formation of spike-like processes, when an endothelial cell is probed and a microneedle pushed into the wall. The spike persists for some time and is firm enough to impede, temporarily, the progress of red cells swept against it by the blood stream. Sanders, Eberth, and Florey (l c.) observed similar projections. However, spontaneous endothelial responses are generally slow and, for the most part, represent an accommodation to changes in pressure. Only when the flow in the vessels ceases may the capillaries decrease in diameter, as their contents drain into neighboring vessels.

In general, the evidence indicates that the capillary wall possesses a variable tone, which Clark and Clark[41] regard as a state of elasticity inherent in living endothelium. The capillaries react to the amount of flow through them. On the other hand, the regulation of the flow

must be ascribed to the vasomotor activity of the muscular vessels, *viz.*, the metarterioles and precapillaries, which feed the flow into the capillaries.

The architecture of the capillary bed, especially the structure of the central channel and the arrangement of its side-branches, is suggestive, regarding the functional activity of the bed. The central channel, being a continuation of the arteriole, maintains a relatively rapid, though decreasing, rate of flow throughout its length. Its wall possesses a gradient of perviousness which progressively increases, so that the outward diffusion of fluid reaches its maximum at the venous end of the channel.[24]

The precapillary offshoots, at their junctions with the metarteriole, and with the proximal a-v capillary channels, take the form of partial twists which progressively diminish, the further distad they come off the central channel. This feature is discussed and illustrated in a recent article by Zweifach.[6] These twists, together with their muscularity, offer resistance to flow from the a-v channel into the precapillaries and, hence, throughout the capillaries. The progressive lessening of the twists conforms with the progressive fall of pressure in the channel. This favors the maintenance, in the true capillaries, of a more or less uniformly low hydrostatic pressure, which facilitates inward diffusion through the walls of the capillaries.

The return of blood from the capillaries into the venous circulation is aided considerably by the acute angled position of the inflowing capillaries, where they join the central a-v capillary channel in the direction of the channel flow. This arrangement, combined with the relatively rapid flow in the a-v channel, should favor the creation of a decrease in pressure at the junction, so as to favor the inflow.

SUMMARY

1. The basic topography of a predominantly nutritive type of capillary bed is presented as a central channel, of which the true capillaries are the side-branches.

A. The different portions of the central channel, in sequence, are: (a) the metarteriole, with typical, but discontinuous, muscle cells imparting vasomotion to the vessel; (b) the proximal portion, with atypical muscle cells; (c) the distal portion of the central channel, with no muscle cells; and (d) the non-muscular venule. The venule acquires muscle cells after it leaves the capillary bed.

B. The precapillaries are the proximal, muscular portions (20–30μ

in extent) of the abrupt offshoots of the muscular portion of the central channel. They act as sphincters, controlling the blood flow in the capillaries.

C. The true capillaries continue from the precapillaries and are also direct (more or less incoming) branches of the distal portion of the central channel and of the non-muscular venule.

2. Vasomotion is a peculiar type of motor activity of the metarteriole and of its precapillaries. It consists of an irregularly recurrent series of dilatations and contractions, at intervals varying from 15 seconds to 3 minutes. It serves to control the extent and distribution of the circulation in the capillary bed. During active metarteriolar vasomotion, the sphincter-like closure of the precapillaries tends to restrict the capillary blood-flow to the central channel. Thus, the central a-v channel carries on the basal work of the bed in ischemia. The true capillaries constitute a reserve bed which comes into action in hyperemia.

3. The endothelial cell possesses a cellular tone which gives to the wall of the capillary a certain degree of elasticity.

4. The true capillary shows passive changes in diameter, as a consequence of the variations of blood-flow through it. When the pressure within the capillaries falls, the endothelial cell tends to lose its expanded state, whereupon its nucleus rounds up and creates a bulge into the lumen of the capillary.

BIBLIOGRAPHY

1. **Dale, H. H., & A. N. Richards**
   1918. J. Physiol. **52**: 110.
2. **Krogh, A.**
   1929. The Anatomy and Physiology of Capillaries. Yale Univ. Press. New Haven.
3. **Bayliss, W. N.**
   1923. The Vaso-Motor System. Longmans Green & Co.
4. **Landis, E. M.**
   1927. Am. J. Physiol. **81**: 124.
5. **Zweifach, B. W.**
   1937. Am. J. Anat. **60**: 473.
6. **Zweifach, B. W.**
   1939. Anat. Rec. **73**: 475.
7. **Chambers, R., & B. W. Zweifach**
   1940. J. Cell. & Comp. Physiol. **15**: 1.
   1944. Am. J. Anat. **75**: 173.
8. **Zimmermann, K. W.**
   1922. Zeit. f. Anat. u. Entw. **68**: 29.

9. **Midsuno, R.**
1930. Beitr. z. path. Anat. u. allg. Path. **84**: 183.
10. **Clark, E. R., & E. L. Clark**
1932. Am. J. Anat. **49**: 441.
11. **Hooker, D. R.**
1920. Am. J. Physiol. **54**: 30.
12. **Lewis, T.**
1927. Blood Vessels of Human Skin and their Responses. Shaw & Sons. London.
13. **Heimberger, H.**
1925. Zeit. f. d. ges. exp. Med. **46**: 519.
14. **Schneider, W.**
1916. Biochem. Zeit. **173**: 111.
15. **Benninghof, A.**
1927. Zeit. f. Zellforsch. u. mikr. Anat. **4**: 125.
16. **Ebbecke, U.**
1923. Pflügers Arch. ges. Physiol. **199**: 197.
17. **Killian, H.**
1925. Arch. f. exp. Path. u. Pharm. **108**: 255.
18. **Fulton, G. P., & B. R. Lutz**
1940. Science **92**: 441.
19. **Schaly, G. A.**
1926. Dissertation **1**. Groningen.
20. **Tannenberg, J.**
1926. Frankf. Zeit. f. Path. **34**: 1.
21. **Klemensciewicz, R.**
1921. Handbuch der biologischen Arbeitsmethoden **5** (4), (1).
22. **Jacoby, W.**
1920. Arch. f. exp. Path. u. Pharm. **86**: 49.
23. **Wolheim, E.**
1927. Klin. Wochensch. **6**: 2134.
24. **Zweifach, B. W.**
1940. Cold Spring Harbor Symp. Quantit. Biol. **8**: 216.
25. **Hagen, W.**
1921. Zeit. f. d. ges. exp. Med. **14**: 364.
26. **Nesterow, A. J.**
1925. Pflügers Arch. ges. Physiol. **209**: 465.
27. **Wyman, L. C., & C. Tum Suden**
1932. Am. J. Physiol. **99**: 285.
28. **Grant, R. R.**
1930. Heart **15**: 257.
29. **Hertzman, A. D.**
1941. Am. J. Physiol. **134**: 59.
30. **Burton, A. C., & R. M. Taylor**
1939. Am. J. Physiol. **126**: 453.
31. **Sandison, J. C.**
1932. Anat. Rec. **54**: 105.
32. **Clark, E. R., E. L. Clark, & R. G. Williams**
1934. Am. J. Anat. **55**: 47.

33. **Wiggers, C. J.**
1942. Physiol. Rev. **22**: 74.
34. **Clark, E. R., & E. L. Clark**
1934. Am. J. Anat. **54**: 299.
35. **Clara, M.**
1939. Die arterio-venösen Anastomosen. Leipzig.
36. **Wright, I. J.**
1932. J. Clin. Invest. **2**: 835.
37. **Krogh, A., & B. Vimtrup**
1932. The capillaries. Cowdry's Special Cytology **1**: 477.
38. **Bensley, R. R., & B. Vimtrup**
1928. Anat. Rec. **39**: 37.
39. **Fields, M. E.**
1935. Skand. Arch. f. Physiol. **72**: 175.
40. **Beecher, H. K.**
1936. Skand. Arch. f. Physiol. **73**: 1.
41. **Clark, E. R., & E. L. Clark**
1935. Am. J. Anat. **57**: 385.
42. **Richards, A. N., & C. F. Schmidt**
1924. Am. J. Physiol. **71**: 178.
1924. Am. J. Med. Sci. **168**: 1.
43. **Crawford, J. H.**
1926. J. Clin. Invest. **2**: 351.
44. **Sanders, A. G., R. H. Ebert, & H. W. Florey**
1940. Quart. J. Exp. Physiol. **30**: 281.
45. **Kahn, R. H., & F. Pollack**
1931. Pflügers Arch. ges. Physiol. **226**: 799.

DISCUSSION OF THE PAPER

Dr. Paul Nicoll (*Indiana University, Bloomington, Indiana*):

In regard to Drs. Chambers' and Zweifach's concept of a central channel through capillary fields, I feel the evidence, to date, is not sufficient to warrant the contention that this is a fundamental organization for all capillary beds. There seems to be no question of the existence of these channels, which serve as a major pathway for blood flow between the arteries and veins, in the mesenteries of some forms. However, until their presence in a diversity of forms and regions is proven, they should be considered as an adaptation of the localities where they are found. The absence of a central channel in the subcutaneous vascular beds of the bat, as determined by direct observation and the weakness of the indirect evidence for their existence in cutaneous tissue elsewhere, would indicate it is not a universally applicable basis for analysis of capillary responses.

The concept of vasomotion has been re-phrased in our final paper, so that any vascular response dependent on muscular action of the smooth muscle elements in its walls is termed *active vasomotion*. Thus, in reference to the active vasomotion observed by Drs. Chambers and Zweifach in their metarterioles, we feel this activity belongs to the type of irregular active vasomotion. This is based on their report of the difference between the duration of the contraction and relaxation phases, and its dependence on an intact nerve supply. It would be interesting to know if the action of the precapillary sphincters in their preparation might not differ from that seen in the metarterioles, and resemble more the rhythmical active vasomotion of the precapillary sphincters and veins, as seen in the subcutaneous

vessels of the bat. This latter type of vasomotion appears to be an inherent response of the smooth muscles themselves; is more regular in the partition of the constrictor and relaxor phases; and does not require intact nervous connections for its appearance.

It is quite possible that the tempo of alternation of constriction and relaxation of irregular active vasomotion might become regular, due to a rhythmical discharge from the vasomotor center or some other factor, but such a regular tempo in irregular active vasomotion is seldom seen. It is necessary to understand, however, that insufficient data have been collected, as yet, to support more than tentatively these subdivisions of active vasomotion.

Dr. Zweifach:

The existence of a structural organization of the capillary bed, built around a framework of thoroughfare muscular channels from which non-muscular true capillary offshoots are distributed, has been established in studies on tissues such as mesentery, omentum, skeletal muscle, intestinal serosa, and subcutaneous tissue in the interdigital web of a variety of mammals and amphibia. In all of these tissues, the same basic vascular pattern was found, except for minor variations introduced by structural peculiarities of different tissues. The vascular pattern in the web of the hind foot of the frog was atypical and resembled the mesh-like capillary network described in the wing of the bat. In the mesentery, omentum and subcutaneous tissues of mammals such as the rat, dog, and cat, many of the central muscular capillary channels lead directly into venules. In skeletal muscle, similar thoroughfare channels exist, but are relatively few in number. The evidence indicates that, irrespective of the precise structural organization of the capillary bed, the integrative feature of the capillary system, which enables it to distribute the blood circulation in accord with the needs of the tissue, resides in the activity of the muscular channels (arterioles, metarterioles, precapillaries) leading into the true capillaries.

Dr. Clark observed that periodic contraction of vessels is widespread in the rabbit's ear.

Dr. Drinker cited the observations of Dr. Florey on the passage of carbon particles through endothelial cells.

Dr. Chambers:

We have never observed carbon particles passing through endothelial cells, such as Florey described. On the other hand, we have frequently seen them and red cells spurt out between endothelial cells through weakened spaces where leucocytes have recently passed. The endothelial cell normally appears in optical section as a mere line. A granule, to pass into a cell, must presumably be taken up by phagocytosis and I have never seen a phagocytic cell expel granular material except by pinching off a part of itself enclosing the particle. This would be a rather uneconomic way for a cell to behave, especially when there are other opportunities for the expulsion of formed bodies, namely, through the intercellular cement, the stiffness of which varies greatly from time to time.

In reply to Dr. Knisely's query, concerning the contractility of purely endothelial capillaries, we have made numerous observations on the effect of the local application of minimum effective concentrations of epinephrin to the capillary bed of the rat mesoappendix. We have never observed a purely endothelial capillary to respond by constricting. Whenever constriction has been seen, it has been of vessels with enveloping muscle cells.

BLOOD CIRCULATION IN THE SUBCUTANEOUS TISSUE OF THE LIVING BAT'S WING[1]

By Paul A. Nicoll and Richard L. Webb

Department of Physiology, School of Medicine, Indiana University, Bloomington, Indiana, and Department of Anatomy, College of Medicine, University of Illinois, Chicago, Illinois.

The wing membrane of the bat offers a field for visualization of the normal subcutaneous circulation of blood and lymph in the living, unanesthetized mammal.[2] In this semitransparent structure, all types of vessels, from arteries through capillaries to veins, may be studied under any magnification desired, without molesting their normal environment The anatomical patterns and their physiological responses are natural and the question of altered environment due to surgical manipulation and the ingrowth of regenerated tissue does not arise. Of the species inhabiting the caves of southern Indiana, *Myotis lucifugus* is the most suitable subject for vascular studies. They are available in unlimited numbers, during their hibernation period between the months of October and May.

For the observations reported here, an unanesthetized bat is slipped into a small metal holder that allows one wing to be extended over a glass plate. The wing is lightly, but firmly, held in the extended position of normal flight by weak spring clips. Contact between the wing surfaces and the glass plate beneath, as well as the coverglass or immersion objective above, is maintained by a film of light petroleum. In experiments where the upper epidermal layer is removed, the exposed tissue is bathed with petroleum and the circulation remains normal for several days. The holder fits into the mechanical stage and permits easy study of extensive vascular fields. Neither surgical procedure nor excessive manipulation is involved in preparing a bat for study. Vital dyes or drugs are introduced between the epidermal layers, by means of a fine pipette, when experiments requiring their use are involved.

An improved holder, better light source, and other technical advances have permitted the use of photomicrographic recording of the vascular

[1] This work was aided by grants from the Graduate School, Indiana University, and the Graduate School, University of Illinois.

reactions in these normal, intact preparations. With cinematographic equipment similar to that described previously,[2] a motion picture record of these subcutaneous vessels and their reactions has been obtained. This was shown at the Conference on Lymph and is available to those who are interested in the subject.

The Vascular Field

For convenience, these studies have been carried out on the wing area between the body wall and the fifth finger. In the average specimen, this region is rectangular, approximately 2½ by 1½ inches, and is supplied by one major artery from the shoulder. The upper and lateral margins of the area receive a few small accessory arteries that anastomose with the main artery or its branches. The venous patterns are identical, in their general distribution, to the arterial ones. From the arterial plexus, smaller vessels arise throughout the field, which, likewise, form interconnected channels or vascular loops. Their size, structural composition, and physiological behavior indicate that they are arteriolar nets. They have a marked capacity for changing the diameter of their lumen and represent the major site of the peripheral resistance. The usual diameter of the lumen in the smaller arterioles within these nets is equal to, or even less than, the diameter of the disk-shaped erythrocyte. From any of these vessels of the arteriolar plexuses, there may arise a vessel whose branches continue as the nonmuscular capillaries. Vessels of this type are the terminal arterioles. They exhibit wide variation in complexity of their subdivisions and possess muscular elements in their walls, for varying lengths along their ramifications. A typical terminal arteriole and adjacent vessels are shown in FIGURE 1. They are sketched to scale, while normal flow is in progress, and regions of active vasomotion are indicated by thickenings in the vascular walls. It is evident that the muscular coat becomes less regular, as the terminal arteriole is followed peripherally. The number of muscle cells found along a branch becomes fewer, the further along the terminal arteriole that the branch arises.

It has not been possible to identify preferential channels through the capillary plexus connecting the terminal arterioles with the coalescing venules. The vascular paths between the arteriolar and venular plexuses in the subcutaneous tissue of the bat follow, in general, the descriptions of Sandison,[3] and Clark and Clark[4] of the subcutaneous patterns in the rabbit ear. No central channel, similar to that described by Chambers and Zweifach[5] in the mesenteries, is found in these sub-

cutaneous fields. At times, especially in the terminal arterioles that receive blood from both directions along the parent arteriolar loop, there appears to be a path of major flow through the capillary and venular plexus. This path, however, is inconsistent, changing into alternate routes through the plexus, with modifications in arteriolar or venular circulation in the adjacent regions. In the majority of the terminal arterioles, no preferential flow path is detectable. Contrary to the usual idea, summarized by Clark,[6] no true arterio-venous anastomoses are found in the bat's wing.

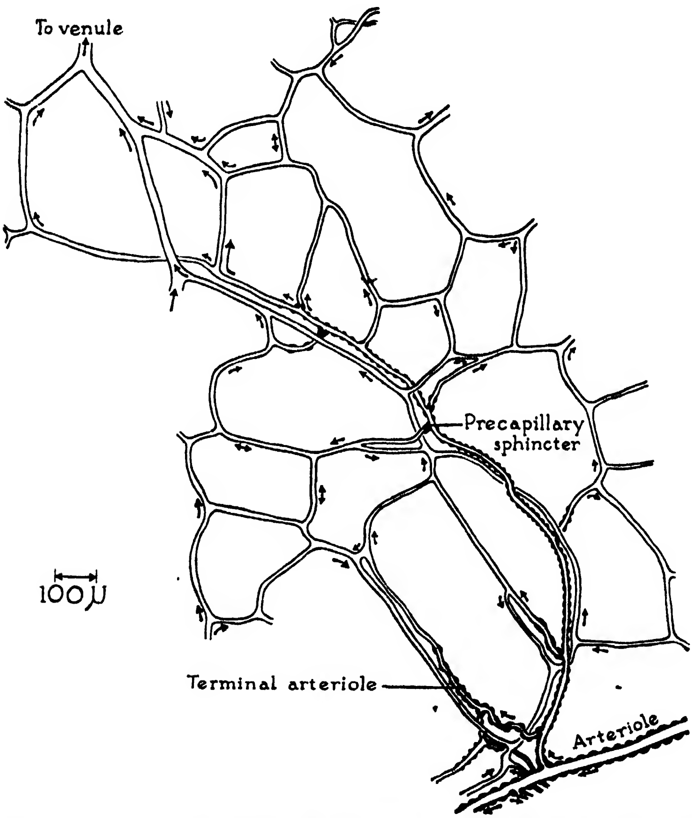


FIGURE 1. Path of blood from arteriole, through terminal arterioles and capillary bed in small area of bat's wing. Heavy irregular lines indicate vessels having muscle cells in their walls. Active vasomotion coincided with the distribution of muscle. This movement was most pronounced in the precapillary sphincters located at the mouths of capillary branches of the long terminal arteriole. The arrows indicate the direction of blood flow in the field during the period of observation. Double-headed arrows show vessels in which the flow was reversed during this time. The length of the arrows indicates the relative flow in each vessel. No preferential channels through such capillary beds could be identified.

Vasomotion

The term, vasomotion, is defined usually as a change in caliber of a blood vessel. Chambers and Zweifach have used it in a more restricted sense, by limiting it to the rhythmical alteration of contraction and dilation in the metarteriole. Preferably, one should indicate a specific kind of change in vascular caliber by the use of suitable adjectives. It is suggested that any caliber change of a blood vessel, produced by contraction or relaxation of muscular components in the walls, be termed *active vasomotion*. Likewise, *passive vasomotion* should refer to changes in caliber produced by alterations of pressure, either internal or external, that are not the result of activity of muscular elements. In this sense, active vasomotion is seen in all blood vessels of the bat's wing that have muscle cells within their walls.

On the arterial side, active vasomotion is manifested by three kinds of vascular response. The maintained contraction of arteries, which is probably a tonus response of their smooth muscle coats, is termed *tonic active vasomotion*. Superimposed upon this tonus reaction is the more rapid contraction or relaxation of vessels, that is dependent on the inflow of nervous impulses. Because of its nature, it is called *irregular active vasomotion*. The final type is a periodic alternation of contraction and relaxation of the vessel, termed *rhythmical active vasomotion*.

That muscular arteries and arterioles at all times are partially constricted, has long been recognized. The marked dilation of these vessels at death, or the dilation of shock and following denervation, indicates clearly the existence of tonic active vasomotion. A manifestation of this action is seen in the bat's wing, in the formation of the "Indian club" enlargements at many bifurcations of arteries and arterioles. Several hours after sectioning the nerve to a limited area, these thickenings of the walls and narrowings of the orifices at the origin of a branching vessel have disappeared. Also, within this area, the average diameter of arteries or arterioles is approximately twice that of the vessels before denervation. Although tonic active vasomotion is certainly due to tonus reactions in the smooth muscle coats, the mechanism of its production, and how the necessary innervation is involved, are not understood.

The outstanding characteristics of tonic active vasomotion are its constancy and sluggishness. They suggest a functional role in a stable system, such as that required of a mechanism for correlating the vas-

cular volume with the total blood volume. If this is the case, it is quite possible that the diverse conditions which precipitate a state of shock may find their final common denominator in their effect on tonic active vasomotion.

The more rapid caliber changes of variable duration and magnitude, exhibited by the arteries and arterioles, comprise irregular active vasomotion. They are the direct result of muscular response to impulses from the vasomotor nerves. The action is the end-organ response of reflexes controlled by the vasomotor center. It ceases immediately upon section of the nerve to the area under observation. The usual response to applied stimuli, such as touching the skin or making of noises, belongs to this type of reaction. Thus, irregular active vasomotion is the mechanism for modifying the peripheral resistance and regulating the pressure gradient within the capillaries.

There is nothing unique in the above description of tonic and irregular active vasomotion. It amounts to a modified restatement of vascular reactions, long recognized and intensively studied. The third type of active vasomotion, however, introduces a concept of peripheral vascular behavior that has recently been brought to attention and is still in the process of crystallization. This rhythmical active vasomotion, or change in the diameter of a vessel, produced by regular alternation of contraction and relaxation of its muscular coat, has been reported casually by several observers. Chambers and Zweifach describe the reaction as seen in the metarterioles of the mesentery. In the subcutaneous vessels of the bat's wings, this reaction is prominent. Arteries, arterioles, terminal arterioles, and the precapillary sphincters all show the reaction. It is the predominant, if not only, type of active vasomotion shown by the veins. In general, the more normal the conditions, the more outstanding is the rhythmical active vasomotion. Anesthesia and hyperemic conditions depress or obliterate the activity, but functional innervation is not a necessary prerequisite for its appearance. The further one proceeds peripherally along the vascular tree from the arteries, the more marked becomes rhythmical active vasomotion. On the arterial side, the best example of the action is seen in the behavior of the precapillary sphincters.

The question arises as to whether this rhythmical activity is the expression of an inherent property of the smooth muscle cells, or one impressed upon them by external factors. The external factors might include rhythmical discharge from the vasomotor center, humeral influences, or physical conditions. Used in this sense, these external fac-

tors should not merely favor the expression of an inherent rhythmical activity of the smooth muscle cell, but actually cause the response. No final conclusion is possible, at present, but observations on the bat's wing give some support to the inherent response hypothesis. They are summarized as follows:

a) Terminal arterioles, precapillary sphincters, and venous vessels exhibit rhythmical active vasomotion after denervation.

b) This type of vasomotion is developed most highly in the precapillary sphincters and in the venous muscular coat, neither of which appears to be under the direct control of vasomotor nerves.

c) It is observed only in the more highly organized arteries and arterioles, that show a moderate degree of tonus and are not responding in an irregular manner.

d) Adjacent vessels, coalescent veins, and even single precapillary sphincters may show rhythmical active vasomotion, completely independent, in rate and magnitude, of their immediate neighbors.

e) Local or systemic application of suitable concentrations of adrenaline enhances the rate and magnitude of rhythmical active vasomotion, at least in the veins, while producing a sustained contraction in the arteries.

The functional significance of rhythmical active vasomotion is obscure. From its predominance in the terminal arterioles and precapillary sphincters, one might conclude that it is a mechanism that allows a minimum volume of blood to serve the metabolic needs of the tissue, at a given pressure gradient across the capillary bed. Thus, all vessels in the area are periodically flushed with fresh blood, without producing a significant change in peripheral resistance. The process is accomplished with a fraction of the total blood volume required to fill the vascular system involved. Such a mechanism might operate normally in relatively inactive tissue like skin or mesentery and could serve during the inactive periods of other tissues. Why it should be so well developed in the subcutaneous veins of the wing membrane, remains a mystery, unless it is a special adaptation that has its physiological basis in the apparent absence of vasomotor control of these structures.

Rhythmical active vasomotion in the veins deserves further mention. It is a striking phenomenon, as was emphasized by its discoverer, Wharton Jones.[7] The action is frequently powerful, with the vascular lumen being reduced to a third or fourth of its resting diameter

at the peak of contraction. The activity moves along the vein in a wave that appears to be organized on an intravalvular segmental basis. This often produces the paradoxical appearance near the end of a segment, in which the movement of the blood is not related to the vascular contraction. Such an observation possibly led to Carrier's[8] unfortunate remark on the absence of a functional significance of the action. There can be no doubt about the importance of this rhythmical active vasomotion in returning blood along the wing veins. A similar segmental organization of the contraction wave is found in the adjacent contracting lymphatics.[2]

The rhythmical, active vasomotion of the veins shows the same fundamental characteristics as that type of activity elsewhere. It occurs usually at a faster tempo, as it is not uncommon to observe a rate of 50 per minute in the veins, while, at the same time, the action on the arterial side is only 12 to 16 per minute. The activity begins along the coalescing venules at the region of the first valves. In the section immediately preceding the first valves, a few isolated bands that are contracting rhythmically may be found. They are isolated rings of smooth muscle and probably represent a single spiral type cell. They indicate the beginning of the smooth muscle coat along the veins, which is complete in the segment behind the first valves.

Capillary Responses

The nature of capillary participation in the peripheral vascular reactions remains obscure. Certainly, much of the difficulty arises from the loose use of the word, *capillary,* in both the physiological and anatomical literature. If it could be agreed, that only the vessels between the arterial and venular systems that lack smooth muscle elements within their walls are capillaries, the problem would be greatly simplified. In this sense, the word, *capillary,* is used throughout this paper and may refer to vessels of different size and permeability. In the subcutaneous fields of the bat's wing, the capillaries originate as continuations of the terminal arterioles and form plexuses of varying intricacy. With the coalescence of these channels, the venules are established in the immediate region of the first valves, with the addition of muscular elements to the walls. The lumen varies along the capillary plexus from approximately the size of the circular erythrocyte, near the arterial end, to two or three times that diameter where they merge into the venules.

The observations made during the last three years on the capilla-

ries in the bat's wing support the position, emphasized by Clark and Clark and sustained by Chambers and Zweifach, that no perivascular cells, of the type designated by Krogh[9] as Rouget cells, exist in this region. The smooth muscle cells, at the transitional points from the terminal arteriole to capillary, end more or less abruptly. Beyond this termination of smooth muscle cells within the walls, no change in diameter of the capillaries, due to activity of any perivascular cells, has been observed.

The question of active endothelial participation in caliber changes of the capillaries is more obscure. Modifications in capillary diameter, not related obviously to blood flow in the immediate field, have been observed. These capillary reactions may involve a considerable length of a vessel or may be restricted to the region of an endothelial nucleus, somewhat as Sanders, Ebert, and Florey[10] have described, but in neither case has it been possible to ascribe the caliber change to either an active contraction or intracellular swelling of the endothelium. The capillary changes are sluggish and might easily result from an elastic recoil of the endothelial wall, due to pressure variations within or without the vessel. Until it is possible to measure accurately the pressures involved and test the capillary behavior with known pressure changes, the causative agent of these capillary reactions must remain in doubt. Regardless of their method of production, the blood flow through a given capillary may be modified by these capillary changes in lumen diameter. In rare cases, even complete stasis may result.

Cytological Components of Vascular Walls

Supravital staining was employed as a means of identification of the structural elements in the fields observed. The dye chosen for general treatment was Methylene blue, Ehrlich vital, National Aniline. This dye, in a concentration of 0.9%, as applied to the tissue, stains a wide variety of tissues and leaves no detectable, irreversible effects in their function. The cells stain progressively, generally, in the following order: mast cells, histiocytes, nerve fibers, fibrocytes, endothelium, skeletal muscle, and smooth muscle. They retain the blue color for varying times, from a few minutes, as in nerve nets, to a day or more, in histiocytes and an occasional muscle cell.

The arteries present the usual histological appearance, having both circular and longitudinal muscle fibers in their tunics. In the smaller branches, endothelial nuclei become evident. As the arterioles are ap-

proached, the circular smooth muscle fibers are arranged in one layer. The longitudinal ones are sparse and eventually are lacking. The characteristic arrangement of smooth muscle cells at right angles to the endothelial nuclei is in evidence.

The terminal arteriole, at its inception, presents the typical arteriolar structure. As the capillary is approached, the muscular layer becomes modified. The first indication of a change is the separation of the closely approximated circular smooth muscle fibers. Areas of bare endothelium become evident. Finally, the cells are spaced at intervals of 20 to 30 microns apart. This transition is abrupt, often involving only 2 or 3 cells. The last of such fibers can be identified in the region of the precapillary sphincters.

Unlike the picture described by Tudor Jones,[11] these muscle cells do not lose their coiled arrangement as the terminal arteriole merges with the capillary. Although separated from their companion cells, they retain their close spiral arrangement about the walls of the vessel, similar to the muscle cells described by Strong.[12] The spiral is formed by 6 to 12 turns. As the diameter of the vessel is approximately that of the circular red cell, a fiber when uncoiled may be 150 microns in length. The nucleus is located at the thicker middle part of the cell and is oriented in its longitudinal axis with the spiral. The presence of such cells at the site of sphincteric vasomotion in the arteriolar capillary junction is indicative of its causative action. The cells just described are little different in appearance from those along the transitional arterioles where rhythmical vasomotion is observed. In these regions, the coils of the spiral cells are closer and the space between them is reduced.

The capillary, as a structural unit of the vascular system, obviously has an end which is connected to an arteriole and one that blends into a venule. Between these extremities, it is devoid of muscular cells. The endothelial nuclei stain with Methylene blue, but the endothelium is not separated from the surrounding areolar connective tissue by fibrous and muscular tunics.

The endothelial nucleus can be identified as a part of the wall of the capillary. Intimately associated with this wall is the adventitial cell. The relationship is prominent, even in unstained preparations, but the reaction of the adventitial cell to the dye is slight. The nucleus stains lightly, and faint protoplasmic processes are seen. The cell seems to be oriented invariably with the long axis of the vessel.

The next cell oriented along, although not in intimate association

with the wall, is the neurilemma cell. The distinguishing feature is its intimate contact with a stained unmyelinated nerve fiber which is usually the distance of a vessel-width from the capillary wall. The staining reaction of the nucleus is marked. Where a neurilemma cell lies on the capillary wall, its identification as a definite entity from the adventitial cell offers no problem, as suggested by Tudor Jones or Michels.[13] The staining reactions of the two cells are different and relation of the neurilemma cell to the nerve fiber is definite.

Other cells of the field, whose distribution is not confined to the region of the capillary and the accompanying nerve, are the mast cells, the histiocytes, and the fibrocytes. The mast cells react in their characteristic manner. The nucleus is stained lightly and the coarse granules react markedly. The histiocytes have a more finely stippled cytoplasm and assume various tortuous shapes. During one phase of the staining reaction, the fibrocytes are in evidence, distributed uniformly over the field.

On the venous side of the capillary beds, the vessels acquire a muscular coat in the region of the first valve and merge imperceptibly into the small vein. The veins acquire the double coat of circular and longitudinal smooth muscle cells, and their histological picture is similar to that of fixed preparations.

Peripheral Nerve Nets and Vascular Innervation

The rapid, powerful constriction of the main arteries and arterioles, when the unanesthetized bat is provoked to struggle, supports the contention that these vessels with well-developed muscular coats receive vasomotor nerves. The absence of such responses in the capillaries throws doubt on their receiving motor innervation.

In all preparations adequately stained with Methylene blue, two nerve nets may be identified easily in any area. They lie deep to each epidermal layer and follow the capillary pattern. It is possible to see many fine terminations of non-medulated free endings ramifying the tissue spaces, but no similar twigs ending on the capillaries have been observed. Frequently, fibers in the nerve nets cross or lie against the capillaries, but no nerve twigs can be traced into the capillary wall in these cases. The adequacy of the staining technique to bring out nerve branches to the capillaries, if they exist, is the observation of motor nerve endings on the thin bands of skeletal muscle. It is possible to follow the branches of the motor nerve on skeletal fibers to their termination in the large motor end plates. These latter struc-

tures stain readily, having a form similar to that described in fixed preparations. It is suggested that the apparent close association of peripheral nerve nets and capillaries is a resultant of lines of stress during development and has no functional significance.

Leucocytes and Capillary Flow

From a comparison of the caliber of small arterioles and capillaries to the diameter of the spherical leucocytes, it is obvious that the latter could plug the vessels unless the internal pressure was able to deform them sufficiently to push them along. This seems to be the case in the small and terminal arterioles where little disturbance of flow, except at the peak of a constrictor phase of vasomotion, is observed. In the capillaries, the low pressure[14] and alternate routes set up ideal conditions for leucocyte plugging. This has been observed repeatedly in many fields, confirming similar observations of Sandison in the rabbit, and is pronounced in the more tortuous capillaries of the small bands of skeletal muscle. Frequently, when the angle of origin is sharp, the leucocyte plugs the entrance of the capillary; while, in other cases, the slight indentations associated with endothelial cell nuclei offer ideal locations.

The removal of these plugs, which are often the determiners of flow in capillary fields, seems to result from two types of mechanisms or a combination of both. In the simplest case, the internal pressure increases, with the resultant dislodgment of the leucocyte. In the second type, the leucocyte slowly fits itself to the vessel lumen or crawls past the obstruction, after which it offers little further resistance and is washed rapidly ahead to a larger vessel, where it returns to its original spherical form. When the leucocyte crawls past an obstruction, such as a protruding endothelial nucleus, the caliber of the vessel is actually widened by the action and may remain so, for some time. It should be emphasized that this type of plugging by leucocytes occurs in normal fields with vigorous flow, and is not related to leucocyte blockage that results from a developing tendency for them to adhere to the vascular walls. This latter action develops several hours after immobilization of a wing and is the first sign of a general migration of leucocytes into the tissue spaces. There is a definite possibility that the opening-up of capillaries in active tissue, stressed by Krogh and others, may depend on changes which greatly reduce or limit the normal plugging action of the leucocytes.

SUMMARY

1. Peripheral blood circulation in the subcutaneous tissue of the normal unanesthetized mammal is observable microscopically in the wing membrane of bats. The vascular components merge gradually from one to another, without sharp differences in structure or functional behavior. Active vasomotion may be seen in all vessels that have demonstrable smooth muscle cells in their walls. Capillaries, without smooth muscle, exhibit passive vasomotion, as shown by sluggish caliber changes of their lumen.

2. The final ramifications on the arterial side are the terminal arterioles, which blend into the capillaries by thinning-out and loss of their muscular coat. Central channels in the capillary nets are not observed.

3. Smooth muscle cells having 6 to 12 coils formed in tight spirals are distributed along the arterioles. Each cell has a nucleus oriented with the spiral and located in a swelling at its mid-portion. The active vasomotion and sphincteric behavior of the arterioles are dependent on the activity of these cells. Adventitial cells play a passive role, if any, in caliber changes of the vessels.

4. Evidence based on supravital staining with Methylene blue, adequate for demonstrating free nerve endings throughout the tissue spaces and the motor end-plates of skeletal muscle, fails to show innervation of capillaries.

5. The role of leucocytes in modifying flow in the normal vascular beds, by occasional plugging of small capillaries, is described.

REFERENCES

2. **Webb, R. L., & P. A. Nicoll**
1944. Anat. Rec. **88**: 351.

3. **Sandison, J. C.**
1932. Anat. Rec. **54**: 105.

4. **Clark, E. R., & E. L. Clark**
1943. Am. J. Anat. **73**: 215.

5. **Chambers, R., & B. W. Zweifach**
1944. Am. J. Anat. **75**: 173.

6. **Clark, E. R.**
1938. Phys. Rev. **18**: 229.

7. **Jones, T. W.**
1852. Phil. Trans. Roy. Soc. London. [illegible]: 131.

8. **Carrier, E. B.**
1926. Observation of living cells in the bat's wing. Physiological Papers: 1–9. Edited by R. Ege, H. C. Hagedon, J. Lindhard, & P. B. Rehberg. Levin and Munkgaard. Copenhagen.

9. **Krogh, A.**
1929. The Anatomy and Physiology of Capillaries: 70. Yale University Press.
10. **Sanders, A. G., R. H. Ebert, & H. W. Florey**
1940. Quart. J. Exp. Physiol. **30**: 281.
11. **Jones, T.**
1936. Am. J. Anat. **58**: 227.
12. **Strong, K. C.**
1938. Anat. Rec. **72**: 151.
13. **Michels, M. A.**
1936. Anat. Rec. **65**: 99.
14. **Hill, L.**
1921. J. Physiol. P. **54**: 144.

DISCUSSION OF THE PAPER

Dr. R. Chambers (*Department of Biology, Washington Square College of Arts and Science, New York University, N. Y.*):

It is interesting that you find so much active vasomotion distributed throughout the minute wing vessels of the bat. You find this to be associated definitely with circularly and spirally arranged muscle cells around vessels, the purely endothelial capillaries being non-contractile.

The much more speedy alternations which you find in the constrictor and dilator phases, than those we have observed in the mesoappendix of the rat, may be at least partly accounted for, by the trauma incidental to the exposure of our material and by our having used an anesthetic which, we have found, not only slows the vasomotion, but may, if the anesthesia is too deep, cause its disappearance.

Is it possible that the replacement of definite metarteriolar thoroughfare-channels from arteriole to venule, by numerous and scattered vasomoting vessels, may be because of the very widespread capillary network, in which a flow must presumably be kept up, under all conditions of a wing, both in flight and at rest?

I would like to suggest that the pulsating venules serve to carry on the venous flow, after the blood has passed through the great spread of capillary vessels with no thoroughfare-channels to carry on the needed pressure. Zweifach has found that pulsating venules in the rat are not infrequent.

Dr. Paul Nicoll:

I believe my comment on Dr. Chambers' paper has some bearing on the active vasomotion, as he describes it, in the rat's mesoappendix. We think the activity he sees is mainly irregular active vasomotion which is also present in the bat's vessels, but in our preparations the rhythmical active vasomotion is more prominent.

In regard to the presence or absence of central channels in capillary beds, it seems to me little more can be said at the present time than that they are, or are not, present in any given region of some animal. It may be possible, no doubt, to draw more significant conclusions as to the factors which have led to their development when our detailed knowledge of the distribution of central channels is more complete.

The real significance of active vasomotion in the subcutaneous veins of the bat's wing, rests in the indication that this behavior is a latent possibility at least in mammalian venous vessels, and its possible role in venous return from any given region of any mammal cannot be excluded until the action has been shown to be absent in that region, under normal conditions.

Dr. Chambers:

Concerning the independent movements of leucocytes within the flowing stream of blood, you have certainly one of the best objects for such a study. There is no

reason why the leucocytes should not exhibit such motility, particularly if the stream is not too rapid to dislodge them from the wall of the vessel to which they adhere. Do you find an inverse relation between rapidity of blood flow and pseudopodial movement of the leucocytes?

While the leucocytes are passing through the wall, we have never seen them form pseudopodia, except on that part of their bodies which is already outside the capillary wall and then only along their margins, close against the outer surface of the wall.

From our experience we would describe diapedesis as follows: The leucocyte, lying over a weak spot between two adjacent endothelial cells, is forced by the blood pressure through the crevice. That portion within the capillary rounds up and progressively diminishes as its contents are pushed through into its expanding outer portion. This lies flattened against the outside of the capillary wall, between the wall and the jelly-like interstitial matrix. The leucocyte, when once completely outside, eventually migrates away by pseudopodial movement.

Dr. Nicoll:

From our studies, we have reason to feel certain that the independent movement of leucocytes within vessels by their adherence to, and amoeboid progression along, the inner wall, is a normal occurrence. It is true, as one would expect, that there is an inverse relationship between this activity and the velocity of flow within the vessel. The action has never been observed normally in the larger arteries or veins, but there is no question that this independent movement of leucocytes is a simple relationship between general physical factors. It is a specific, positive reaction in which a given leucocyte will move along the inner wall by means of visible extended pseudopodia and typical amoeboid responses. The reacting cell may move in any direction, either up or down stream, and usually does so along an irregular spiral path. This latter action has suggested that the movement follows intercellular cement lines, but there is no proof that such is the case. We have never seen leucocytic diapedesis occur in a normal field, where the cell involved did not first explore by amoeboid movement the vascular wall for some distance before penetration commenced. It is our impression that the penetration is, for the most part, due to positive action by the cell.

An additional factor in favor of the active participation of the leucocyte in diapedesis is the site where it occurs in the normal field. We have only observed the reaction in the coalescing capillaries where the lumen diameter is definitely enlarged over the initial components of the capillary bed on the arterial side. These are still true capillaries, as they have no muscular elements in their walls. In the smaller capillaries, as they first emerge from the terminal arterioles, the lumen is usually about the size of the circular erythrocyte, and the white cells are definitely deformed as they pass through them. It is here that leucocyte plugging most frequently occurs and the most ideal condition for pressure extrusion of a leucocyte would seem to be present. Nevertheless, diapedesis has never been observed in these vessels, even when positive amoeboid activity was observed as a leucocyte crawled past an obstruction. There is need, of course, for additional studies to clarify the true nature of normal diapedesis.

Dr. Chambers:

I have this comment on Dr. Nicoll's interesting observations. It is quite possible and entirely likely that normal diapedesis involves active participation on the part of the leucocyte. Two other features which may also be borne in mind are the driving force of the blood pressure, when once the leucocyte has lodged itself in an interendothelial space, and, second, a reaction of the general region around the capillary which induces a softening or weakening of the interendothelial cement, thus facilitating the egress of the leucocyte. I mention the latter as a case which predominates in inflamed regions where diapedesis becomes excessive to the extent that not only leucocytes, but also erythrocytes, become extravasated without actual rupture of the vessel wall.

Dr. Bret Ratner (*New York University, N. Y.*):

The study of terminal vessels presents intriguing possibilities for an explanation of problems related to the allergic phenomenon. Since vascular changes do

occur in the hypersensitive manifestations of wheal formation, the Arthus phenomenon and periarteritis nodosa, it seemed possible that an explanation for wheal formation might be found in the behavior of terminal vessels. A theory, which I called "the arteriolar spasm theory of wheal formation," was presented in 1943 (Allergy, Anaphylaxis and Immunotherapy. Williams & Wilkins, Baltimore, Md.). It was formulated on the basis of an analysis of the anatomical and physiological characteristics of the terminal vessels.

My reasoning was as follows:

1. There is a preponderance of smooth muscle in the arteriolar side of the capillary loop and an absence of smooth muscle on the venous side.
2. The antigen-antibody concept of the mechanism of allergic phenomena postulates that allergic antibodies are anchored, in hypersensitive tissues, to smooth muscle cells.
3. Specific antigen entering the circulation contacts these specific sessile antibodies present in the smooth muscle cells and combines with them.
4. The irritation of the muscle cells, arising from the combination of antigen and antibody, may, as a result, produce a spasm of the arteriolar side of the capillary loop.
5. The spasm would then cause a partial or complete occlusion only of those vessels containing smooth muscle. It would not affect the terminal vessels devoid of muscle fibers. Thus, solely the arteriolar vessels would be affected.
6. Such constriction of the arteriolar side might conceivably result in a sudden drop in arteriolar pressure.
7. The blood pressure on the venous side of the capillary loop would, as a result, be higher than that on the arteriolar side, while the occlusion of the latter persisted.
8. If slight rises in venous pressure have a tendency to produce edema, it seems reasonable to assume that herein might lie the explanation for an increased transudation of plasma into the lymph spaces through the more highly permeable venous side of the loop, which, in turn, would produce the edematous white swelling characteristic of the wheal. In turn, the edematous focus, if great enough, would obliterate all the vessels in its immediate area by sheer external pressure.
9. The physiological disturbance brought about in this way might cause a dilatation of collateral arterioles and venules, accounting for the surrounding erythema seen in marked wheal formation.

The arteriolar spasm theory relates only to the *initiating phase* and does not attempt to offer an explanation for the secondary by-effects resulting from the disturbed physiology. Histamine-like substances and products of anoxia might well enter at this point to play a complementary role and cause further dilatation There are those who believe histamine is the initiator of the allergic phenomenon, but these advocates leave out of consideration *localization* of the allergic phenomenon. To my mind, localization is explained by the anchoring of specific antibodies in the smooth muscle cells of the various organs; reactions occur at those sites containing the preponderance of smooth muscle cells. At least, anatomically, this would seem more feasible than assuming that a humoral substance, such as histamine, in some unknown way brings about a localization of the allergic phenomenon.

We accept bronchiolar spasm in the guinea pig, hepatic venule spasm in the dog, pulmonary artery spasm in the rabbit, and gastric spasm in the calf for the explanation of anaphylactic syndromes in these various species. Arteriolar spasm has been demonstrated by Abell and Schenck in the ears of rabbits in anaphylaxis. In view of these demonstrated facts and in the light of the analysis of the characteristics of the terminal vessels, it seems to me to be reasonable to propose the theory of a spasm of the terminal arteriole, caused by the combination of antigen and antibody, as the *initiating* factor in wheal formation.

CAPILLARY PERMEABILITY AND THE FACTORS AFFECTING THE COMPOSITION OF CAPILLARY FILTRATE

By Eugene M. Landis

Department of Physiology, Harvard University Medical School, Boston, Mass.

In a conference dealing with lymph and lymphatics, the appropriateness of discussing the permeability of the capillary wall lies in the fact that through this membrane passes the capillary filtrate, which is the ultimate raw material from which lymph arises. Under most conditions, the proved and conjectured relations between capillary filtrate, capillary absorbate, tissue fluid and fully elaborated lymph are so close as to defy exact definition of boundaries. Moreover, changes of colloid osmotic pressure, hydrostatic pressure, or tissue pressure will affect immediately the volume and composition of all four fluids. While changes in blood and lymph can be analyzed accurately, the simultaneously occurring changes in the more abstract intermediary fluids are still largely hypothetical and, for the most part, based on indirect evidence.

Even in the smallest vessels, the blood is separated from the tissues, tissue fluid, and lymph by a single layer of endothelial cells which form the walls of the capillaries. The fundamental properties of this membrane, in conjunction with the osmotic and hydrostatic pressures exerted by the blood and tissue fluid, will determine whether tissue fluid volume will be greater or less than normal, whether filtration or absorption will predominate in a given area, whether lymph volume will be high or low and whether this lymph will contain much or little protein.

It is generally agreed that the capillary wall permits water and many dissolved substances to pass more easily than do the surface membranes of tissue cells or of certain egg cells. The capillary wall appears to be about 3,000 times more permeable than these cellular membranes, when studied quantitatively under comparable conditions.[1] The "conditions" of these observations merit definition, because, in discussion of capillary permeability, they are all-important, though often neglected. In scientific and clinical literature, the term, *capillary permeability,* has

often been used very loosely. It has been confused and used interchangeably with *capillary fragility,* though the latter, in the strictest sense, refers to the passage of red cells. The simple observation that more of a substance or more of a fluid passes through the capillary wall has been presented as direct evidence of increased capillary permeability.

Depending upon conditions, passage of a substance in increased amount does not always mean increased permeability, nor does its failure to pass mean impermeability. It is easy to demonstrate *in vivo* that the capillary wall may be permeable to protein and unable to retain any fluid, while external conditions prevent this change from becoming grossly obvious. Conversely, more fluid may pass through a capillary wall, not because its permeability is greater, but merely because the pressure within the capillary has been increased.

Hence, in defining the permeability of any membrane, living or dead, normal or abnormal, it is necessary to know at the very least: (a) the unit of volume or mass of substance passing through (b) unit area, in (c) unit time, under the influence of (d) unit hydrostatic or unit osmotic pressure. For complete physical accuracy, it is necessary also to know (e) the exact thickness of the membrane, so that the passage of substances can be defined per unit thickness, as well. In the case of the capillary wall, however, most interest attaches to the intact membrane, as it occurs in the living organism; for that reason, the term, *capillary permeability,* by common usage is applied to average total thickness.

What can be said, at present, concerning the physical structure of the membrane upon which capillary permeability depends? While the single capillary is the smallest unit of interchange between blood and tissue, it must be emphasized that these so-called units are by no means identical, and that any "average" figure includes a wide range of normal variation, as well as great individuality on the part of each capillary. Single capillaries differ in diameter, length, and the nature of their connections with the smallest arterioles and venules. Even the wall itself is not homogenous, because it consists of thin, plate-like endothelial cells placed edge to edge to produce a mosaic type of membrane about 1 micron in thickness, which, except for localized bulges, contains the single nucleus of each endothelial cell. The apposed edges of the endothelial cells are bridged, or made tight, by an intercellular substance or cement which will be discussed in more detail, later. This is about as far as the microscope can assist directly in describing this

membrane, because the pathways through which dissolved substances and water pass from the blood to the tissues are beyond the powers of any existing ultramicroscope and, for the present, can only be described by comparing the sizes of those molecules which pass and those which are retained.

By this approach, however, the permeability of the capillary wall must still be estimated indirectly, because an exact and direct definition would be possible only if a true capillary filtrate could be obtained under completely normal conditions and then analyzed for comparison with the blood plasma passing along the lumen of the capillary in question. Samples of truly normal capillary filtrate have not, so far, been obtained, because of the minute volume of this filtrate under average conditions. Moreover, the absorptive function of the capillaries in the normal tissue at normal pressures begins modifying this capillary filtrate, at the moment it becomes extravascular fluid. Information concerning the passage of protein through the capillary wall has necessarily been obtained chiefly from analyses: (a) of blood plasma during venous congestion; (b) of edema fluid collected either from normal subjects during venous congestion or from patients with various clinical conditions; and (c) of lymph collected from small or large lymphatic ducts during rest, activity, or congestion.

It is generally agreed that water passes most easily and that substances such as urea, potassium, sodium, chloride, and nitrate pass almost as easily as water. This would be expected from their low molecular weight and small dimensions. Moreover, by injecting heavy water intravenously, Hevesy and Jacobsen[2] observed in the rabbit that it required only half a minute for water to pass through the capillary wall and penetrate through all the extravascular space. The diffusion of heavy water through a space of 20 micra is almost complete in 0.1 second. Flexner, Gellhorn, and Merrell[3, 4] also observed that, in each minute, 73 per cent of heavy water and 60 per cent of radioactive sodium in blood pass to and fro through the capillary wall between blood and tissue fluid. Calculated on the basis of plasma only, water is exchanged twice as rapidly as sodium.[4]

In the next group of substances are calcium, magnesium, and glucose, but even these pass relatively easily, so that equilibrium of concentration is reached with only moderate delay and their osmotic effects on fluid movement, though definite, are comparatively slight and transitory, as Starling first pointed out clearly. Studies of the most varied

character have failed to produce any evidence that the capillary wall acts otherwise than as a simple filter; it has not, so far, been detected in the act of secreting any substance.

As we come to larger molecules, however, they are retained for a longer time after intravenous injection. Gelatin and isinglass, though fairly satisfactory as temporary substitutes for plasma proteins, are far from ideal, because they pass through the capillary wall into the tissue fluid and into the urine at rates as high as 50 per cent in one-half hour, as Little and Dameron[5] have shown. Their equatorial diameters, from recent measurements by Cohn and his co-workers, are 18 and 16 Ångstrom units, respectively, while the plasma proteins which are retained have equatorial diameters ranging from 32 to 38 Ångstrom units.[6]

At one time, there was considerable discussion as to how efficient the capillary walls in the extremities actually are in retaining plasma proteins. Taking together all the evidence from many sources, it can be concluded that the capillary endothelium in the extremities is, on the average, about 90 to 95 per cent efficient as a protein-retaining membrane, and that the original normal capillary filtrate contains from 0.2 to 0.5 per cent protein. The capillaries of the liver and intestines are, of course, much more permeable, in that roughly 50 per cent of the plasma proteins escape.

From the beginning of these studies, it was perplexing to find that normal capillary filtrates and normal lymph from the extremities contained not only albumin with a molecular weight of 69,000, but also some globulin with a molecular weight of 160,000, and even minute amounts of fibrinogen with a molecular weight of 500,000. While the passage of fibrinogen and of globulin was definitely less than that of albumin, it was still difficult to reconcile this passage with the fact that their molecular weights were, respectively, almost seven and three times greater than that of albumin.

An explanation for this apparent anomaly has been provided by recent studies of physical chemists, such as those of Dr. Cohn and his group.[6] The equatorial diameters of the plasma proteins all range from 33 to 38 Ångstrom units. The differences in molecular weight are associated with differences in molecular length, not with change in equatorial diameter. Therefore, other things being equal, the albumin molecule can be expected by chance to pass more readily than the longer fibrinogen molecules; though, from the standpoint of equatorial diameter alone, globulin and fibrinogen should also pass, when properly placed with reference to any pore or space which permits albumin to pass.

From this and other evidence, it appears that the passage of protein, to the amount of roughly 5 per cent of its concentration in plasma, indicates the presence in the capillary wall of a few pores having an effective diameter of at least 38 Ångstrom units. Estimates arrived at by other means have indicated that pore size may vary from 7 to 20 Ångstrom units, definitely less than the narrowest diameter of the plasma proteins, but still considerably larger than the diameter of the sodium and chloride ions which pass quite freely. Therefore, as a working concept, one may imagine the capillary endothelium to consist of a meshwork with pores of many sizes, of which a few must have diameters of 38 Ångstrom units or possibly more.

While proteins are being retained, or are passing with difficulty, the most diffusible substances, such as urea, salts, glucose, and amino acids, are able to move back and forth quite freely, or at least moderately freely, through the smaller and larger pores by simple diffusion. Fortunately, for the sake of rapid and uniform interchange, this diffusion need not follow the current of water, during either filtration or absorption. Rous and his co-workers,[7] for instance, have shown that certain dyestuffs can diffuse outward through the capillary wall, even while fluid is moving inward by absorption. They observed, also, that the venous capillaries and small venules permit dyes to pass more rapidly, and have described a gradient of permeability by which the diffusion of small molecules becomes more rapid as the venules are approached. This slight difference in diffusion of dyes, maintained below the range at which proteins would pass, should speed and equalize the supplying of tissue cells with substances such as glucose and amino acids. Since permeability does not increase enough to permit large amounts of protein to pass, it should not diminish the absorption of water. It may even expedite that absorption, leading to an additional factor of safety against the appearance of edema. Fluid balance is maintained and this slight increase in permeability to dyes of moderate molecular size in no way invalidates the Starling hypothesis.

The observations of Rous and his co-workers[7] on passage of dyes are interesting, also, because of the ingenuity with which conditions were varied to show the relation between the passage of dye and known changes in capillary permeability, capillary blood pressure and capillary blood flow. Unless used carefully in this way, dyes are treacherous agents for the measurement of capillary permeability. Often impure, loose, chemical combinations, they may be more or less adsorbed to the plasma proteins. Since they bear a charge, their passage through the

capillary wall may be modified to unknown degree, in some fashion related to what is known about their passage through capillaries made of collodion. Storage of dyes by vital staining of tissues outside the vessel may reduce free concentration and maintain a diffusion gradient which would soon be eliminated, were staining absent. For these reasons, alleged changes in permeability, based simply upon dye passage, are surrounded with doubts, unless elaborate control observations are supplied to prove that special factors or artifacts have been eliminated completely.

Studies with dyes have, however, suggested visually the possible temporary importance to fluid interchange, of solutes with molecular dimensions and diffusibilities between those of sodium, potassium, or chloride and those of the proteins. It is known that hypertonic sucrose, glucose, or sodium chloride produce temporary shifts of fluid across the capillary wall. Moreover, in attempting to explain the losses of fluid from the blood, which have been demonstrated during muscular exercise, it became obvious that the rise in capillary blood pressure in exercising muscle could not explain more than 10 per cent of the rapid capillary filtration estimated indirectly from comparisons of arterial blood with venous blood leaving the muscle. To explain this discrepancy, it has been suggested[1] that metabolic products produce not only the well-known vasodilatation with local increase of blood flow and capillary pressure, but, since they are also osmotically active, they can also cause fluid to move toward active tissue, by raising temporarily and conspicuously the total osmotic pressure of the extravascular fluid. In favor of this concept is the rise in vapor pressure of the blood during exercise, as observed by Margaria.[8] The osmotic pressure rose from the equivalent of 0.945 gm. of NaCl per 100 gm. of water to the equivalent of 1.048 gm. of NaCl per 100 gm. of water. Since these osmotically active substances diffuse from muscle to extravascular fluid and from there into the blood or lymph, it is logical to suppose that fluid will be withdrawn from the blood stream, until diffusion restores the osmotic equilibrium which usually obtains between these diffusible constituents of blood and tissue fluid. This would increase the volume of capillary filtrate during exercise, in conjunction with the changes in capillary pressure and permeability which occur also in exercise.[9, 10]

Keys[11] has gone a step further and has called attention to the potentially harmful magnitude of the fluid shifts that could be produced by these osmotically active metabolites. He found that the rise in the total osmotic pressure of the blood during exercise could not be ac-

counted for entirely on the basis of new ions from the tissues. He believes, therefore, that the plasma crystalloids cannot filter out of the blood stream as rapidly as the water which is drawn from the blood toward the muscles by these metabolites. Thus, it would appear that, from the standpoint of very rapid filtration, the capillary wall may be only relatively permeable, even to monovalent ions, just as it is only relatively impermeable to the plasma protein molecules. If this be true, a slight but definite lag in the passage of these crystalloids would provide a powerful force to prevent undue reduction or increase of blood volume by sudden shifts in osmotic pressure on either side of the capillary wall, achieving, thereby, an "osmotic buffering." This concept is difficult to reconcile with the presence of pores which all have diameters of 16 Ångstroms or more. It might be reconciled, however, with the assumption that small pores predominate and that only a very few pores have diameters of 16 to 38 Ångstroms. It should be emphasized, also, that this purely mechanical concept may be far too simple and that other forces, as yet unknown, may explain more easily and logically those facts which are extremely difficult to reconcile with present information concerning the minutiae of capillary permeability.

Next to the absolute permeability of the capillary wall itself, capillary blood pressure is of the greatest consequence. This is not because capillary blood pressure *per se* affects permeability, but merely because a high capillary pressure may increase many-fold the visible effects of a given increase in permeability. Conversely, a very low capillary blood pressure may almost obliterate the visible effects of a marked increase in capillary permeability. On this account, accurate studies of capillary permeability in the intact organism are difficult to interpret with assurance, unless one can simultaneously exclude changes in capillary blood pressure, blood flow, and the area of capillary wall available for interchange. All of these factors are changing within wide limits, under normal conditions, and variability is apt to be even greater under abnormal conditions.

Micro-injection measurements of fluid movement and capillary pressure in individual capillaries have made it possible to express these filtrations and absorptions in terms of cubic micra of fluid being absorbed, or filtered, per square micron of capillary wall per second. These rates of fluid movement can then be charted against capillary blood pressure; thus, the meaning of the Starling hypothesis and its relation to capillary permeability become clear. For the normal capillaries of the frog's mesentery, when capillary pressure is above the col-

loid osmotic pressure of frog's blood, that is above 8 to 12 cm. of water, filtration occurs at an increasing rate. When capillary pressure is below the colloid osmotic pressure of the blood, absorption occurs, also at an increasing rate. The capillary wall acts as an inert, that is, non-secreting, relatively protein-tight filter.

This relation depends, however, upon the colloid osmotic pressure of the plasma proteins, as well as on capillary permeability. Recently, in connection with some studies on the toxin of *Clostridium oedematiens* now being made by Dr. Brown in my laboratory, control observations on the mesenteric capillaries of presumably normal winter frogs indicated that both filtration and equilibrium tended to occur at lower capillary pressures than usual. No clear absorption could be detected, even at low capillary pressures. The specific gravity of the blood plasma from these winter frogs indicated that the plasma protein concentration was so low that absorption could not be expected at existing capillary pressures. The low protein concentration may be due to the low food intake of frogs during the winter months.

The introduction of pure bovine serum albumin into the lymph sac eventually elevated the specific gravity of the blood plasma, and absorption then appeared. These studies will be continued, using spring and summer frogs to determine whether this is the true explanation of what might be assumed, at first sight, to be a seasonal difference of capillary permeability. They indicate, however, the care that must be exercised in drawing conclusions concerning permeability in a system as complicated as the capillary network.

The relation between capillary pressure and fluid movement is quite different when the capillaries have been injured by alcohol or mercuric chloride. No absorption and no equilibrium are seen at any pressure, and the slope of the line relating pressure and filtration indicates that permeability to fluid is increased from seven to nine times above normal. The absence of absorption at any pressure indicates that the effective colloid osmotic pressure of the plasma proteins is reduced almost to zero and that most, if not all, the plasma proteins are able to escape freely with the fluid. Under these circumstances, loss of plasma into the tissue spaces is to be expected, and a capillary filtrate with a high concentration of protein will accumulate rapidly, particularly if capillary blood pressure is high. Equally important, however, from the practical standpoint, is the fact that a very low capillary pressure may mask almost completely the effect of increased permeability.

This has been demonstrated most clearly by Lewis,[12] in the case of

the histamine wheal. When needle punctures are made into the skin through a droplet of histamine, Lewis has demonstrated an increase in permeability, due to the direct effect of histamine on the adjacent capillaries. Local edema, in the form of a wheal, appears. Fluid drawn from this wheal has a very high content of protein, usually 4 to 5 per cent. This result is observed, however, only if circulation is free. Temporary, complete obstruction of blood flow, as by inflating a pneumatic cuff to a pressure above systolic level, delays the appearance of the wheal until blood flow is resumed. Under these conditions, the change in capillary permeability is present as before, but compression of the artery prevents the appearance of local edema, partly, no doubt, because filtration pressure is extremely low and partly because there is no renewed supply of plasma from which capillary filtrate can be formed in large volume.

The effects of injury on capillary permeability can be shown more directly, though less quantitatively, by micro-injection.[13] When a capillary is filled with colloidal dye and then injured locally by compression with a microscopic glass rod, the dye passes immediately and diffusely through the injured area, while it is still held back by adjacent areas of normal capillary endothelium. Even the injured area retains India ink particles and erythrocytes, while the plasma with its protein passes through freely. From similar observations with starch and other substances, Krogh[14] came to the conclusion that the pores of the chemically injured capillary wall have diameters ranging from 5 millimicra to not more than 200 millimicra. This change in pore size is far more than enough to explain the seven- to nine-fold increase in fluid filtration and the free passage of protein indicated by micro-injection measurements of fluid filtration.

It seems to be a general rule that injury of tissues by any means, thermal, chemical, or mechanical, is always accompanied by prompt leakage into the tissue spaces of a large volume of capillary filtrate, which is rich in protein. These effects are immediate and become clearly obvious within a few seconds or minutes after the insult. The observations on *oedematiens* toxin, referred to previously, suggested the existence of another type of increase in permeability, which requires several hours to develop. A lethal dose of toxin injected into the lymph sac of the frog has no measurable effect on filtration for 2 to 3 hours. From 3 to 4 hours, the sticking of leucocytes, the unusual number of leucocytes in the capillary network, and sluggish flow are signs of early damage, but a few quantitative measurements during

this period have, so far, shown little change in fluid movement. It requires from 4 to 6 hours for unmistakable and general increase in permeability and true capillary stasis to develop. At the end of this period, the heart and lymph hearts are still beating.

The direct application of very strong solutions of the toxin to the mesentery also has no effect over 2 or 3 hours. The long latent period between the injection of the toxin and the manifest appearance of injury suggests that this toxin damages the capillary wall by a different mechanism from that which is involved in the immediate stasis produced by mechanical trauma, heat, or chemicals.

The injury produced by alcohol, mercuric chloride, histamine, burns, and inflammation is clear cut and extreme. Less drastic and more physiologic types of stress produce less marked changes in capillary permeability. The accumulation of carbon dioxide has no measurable effect on capillary permeability, nor do changes in hydrogen ion concentration, when tested within physiologic limits. When a pH of 4 is reached, permeability increases, but this is clearly due to gross damage far outside the physiologic range.

Oxygen lack, if mild, has little effect, but prolonged and sustained anoxia of the frog's mesentery increases capillary permeability both to plasma and to fluid.[15] This effect is reversible with brief periods of oxygen lack, but irreversible when oxygen lack persists for more than a few minutes. It must be emphasized, however, that the grade of anoxia which produces this change in capillary permeability is very severe. Results obtained under these extreme conditions cannot be carried over without further test to the milder anoxias that occur, for example, in cardiac failure or in transient vasoconstriction. In some studies, edema fluids collected from patients in cardiac failure contain more protein than normal, indicating an increase in capillary permeability due to anoxia; but, in most instances, the protein content is normal, that is, less than 0.5 per cent.[16] There can, however, be no doubt of the gross capillary damage observed in the extreme and prolonged anoxia of total arterial occlusion. Further work is required to determine, with greater accuracy, what oxygen tension and what duration at each tension is necessary to produce, uniformly, an increase in capillary permeability under reasonably physiological conditions.

The margin of safety may actually be larger than is usually suspected. In venous congestion of the human forearm, the loss of fluid becomes greater as venous and capillary pressures are increased.[17] Average fluid filtration ranges from 1.3 cc., from each 100 cc. of blood

at 20 mm. Hg, to 15 cc., from 100 cc. of blood at 80 mm. Hg. Protein loss is minimal and within the normal limits of 0.2 to 0.5 per cent, up to a venous pressure of 60 mm. Hg. Above that level, the capillary filtrate, by a rather indirect and not too accurate calculation, appears to have an average protein content of 1.2 per cent, suggesting that marked congestion over 30 minutes has significantly increased capillary permeability. Both albumin and globulin pass through the capillary wall, but the loss of albumin is greater. However, it is conceivable, here, that the passage of protein may have been produced in part mechanically, because the capillary walls are all tightly stretched by abnormally high and sustained intravascular pressure.

Similar difficulties are met with, in attempting to decide whether or not a given substance reduces abnormal capillary permeability. The situation is, perhaps, simplest in perfused tissues. Here serum and red cells prevent edema formation, because of their colloid osmotic pressure and oxygen carrying capacity, respectively. However, in some circumstances at least, both have beneficial effects which seem to be greater than expected from these sources alone. Platelet fragments[18] or even erythrocytes and indifferent particulate matter[19] in the perfusion fluid may reduce permeability by mechanical plugging of endothelial pores.

In the case of epinephrine, it seems fair to state that its action in diminishing the local edema of urticaria is completely unexplained In very large doses, as Rigdon suggests,[20] it may prevent inflammatory edema by producing arteriolar constriction. Very small doses tend to produce dilatation in some areas, but, in any case, a direct effect on capillary permeability has not been demonstrated conclusively.

The history of work on pituitrin and calcium is checkered, but in the intact organism they, too, seem to have little effect except as they may produce local vasoconstriction, thereby reducing blood flow, filtering area, and capillary pressure. Impressions vary also concerning the possible effect of extracts of adrenal cortex, but a survey of recent work makes it also very doubtful whether they affect capillaries rendered permeable by trauma.

Temperature unquestionably modifies fluid movement,[21] and it is tempting to assume, therefore, that capillary permeability is increased by mild heat and decreased by mild cold. Yet the protein content of lymph[22] and the passage of dyes[23] indicate that no significant change in permeability appears until injurious grades of heat are applied. As

soon as burning occurs, copious edema fluid and lymph with high protein content appear. The same is true of injurious grades of cold, as, for example, in freezing.

To return again to the physiological range, it is possible to measure accurately the increase in extravascular fluid in the human forearm produced by venous congestion over set periods. Filtration of 44° is twice that observed at 14° C., but it is known that the vessels are dilated by heat and the area available for filtration is correspondingly greater. It seems likely that heat in the range of 15 to 44° C. modifies fluid movement by changes in filtering area, rather than by modification of capillary permeability. The accumulation of edema fluid at temperatures of 10° C. or less, as described by Lewis,[24] suggests injury from cold, but the evidence is not, as yet, complete. Lange,[25] on the other hand, has recently described a reduction in the rate at which the dye fluorescein diffuses into cooled tissues. This, however, may be due to a reduced area available for diffusion, as well as for filtration, rather than to a decrease in permeability.*

So far, the capillary wall has been considered as if it were a simple filter and as if a few molecules of protein represented the largest bodies that can pass through. Yet, the earliest observations of pathologists showed that leucocytes pass in large numbers through the capillary wall into inflamed tissues. Field and Drinker[26] have demonstrated that erythrocytes appear in the lymph collected during venous congestion and during exercise. Graphite particles and pneumococci appear in lymph, within 10 to 20 minutes after they are injected intravenously.

That the capillary wall is not a uniformly resistant membrane, can be shown very easily again by micro-injection. If a frog's capillary be blocked at both ends, and then an India ink solution be injected, by means of a micro-pipette, into this closed endothelial sac, it is found that the capillary wall retains the India ink only up to a certain pressure. As pressure is increased to between 50 and 80 mm. Hg, the India ink will suddenly spurt through a few isolated spots, though no tears of the endothelium can be seen. When pressure is reduced, these particles remain as localized, dense collections just outside the capillary wall. Obviously, the endothelium is not uniformly resistant to distention,

* Recent observations, made with the "pressure plethysmograph"[21] by Brown, Wise, and Wheeler,[22] have shown that exposure to cold (14° and 4° C.) produces filtration and increases the volume of extravascular fluid, even in the uncongested forearm. They have shown also that cold increases the passage of protein, and, therefore, that capillary permeability is increased, not decreased as Lange[25] was led to believe from the slower passage of the dye, fluorescein, into cold

but contains certain weak areas through which particulate matter can be forced. In capillaries not too damaged by micro-injection itself, flow may be resumed normally and without stasis, after India ink has thus been forced through.

If the endothelial layer can thus be forced open and then close again without subsequent loss of protein, it appears that two elements must be considered, the endothelial cell and the intercellular cement. This cement appears to be a calcium proteinate which disintegrates in perfused preparations when calcium is absent.[27] Simultaneously, during perfusion with calcium-poor fluid, red cells and fluid may escape. Chambers and Zweifach have suggested that this thin layer of cement controls general capillary permeability. With the evidence at hand, however, it hardly seems justifiable to locate in this thin intercellular layer more than a control of capillary fragility, that is, the function of bridging the endothelial cells, even though, under certain conditions, it can permit particulate matter such as carbon particles and erythrocytes to pass.

The dissociation between increased capillary permeability and increased capillary fragility is often quite striking. The abundant edema fluid of urticarial lesions and of most inflammations contains few erythrocytes, but a great deal of protein, representing, therefore, a relatively pure increase in permeability. On the other hand, the punctate or gross hemorrhagic lesions seen in the purpuras consist chiefly of erythrocytes, with little or no edema fluid. When free passage of erythrocytes is found, it would seem to indicate a grosser loss of endothelial integrity than is found with simple increased permeability. On this basis, capillary fragility and capillary permeability refer to different and distinctive properties of the vessel wall. Even in the case of the less commonly occurring hemorrhagic wheals, the distinction is still useful, because the passage of red cells in large numbers, in addition to edema fluid, indicates a loss of integrity which is more extensive than the uncomplicated change of permeability which produces the ordinary wheal with accumulation of extravascular fluid rich in protein only.

Finally, to summarize, the conditions included in a strict definition of permeability are such that this term must be used carefully to avoid misunderstanding. Before ascribing the greater or lesser passage of a given substance to a change in permeability, the conditions under which that passage occurred must be scrutinized. Conclusions concerning capillary permeability are sometimes arrived at, without considering

the many simple physical forces which are concerned in the movement of dissolved substances through the capillary wall. It must be admitted immediately that the most careful control of all known physical factors fails to explain some of the commoner manifestations of abnormal capillary function. Yet, it seems quite clear that, in further studies of capillary permeability, adequate control of known forces will make our concept of capillary permeability more quantitative, and place in sharper relief other factors still unrecognized at present.

BIBLIOGRAPHY

1. **Landis, E. M.**
   1934. Physiol. Rev. **14**: 404.
2. **Hevesy, G., & C. F. Jacobsen**
   1940. Acta Physiol. Scandinav. **1**: 11.
3. **Flexner, L. B., A. Gellhorn, & M. Merrell**
   1942. J. Biol. Chem. **144**: 35.
4. **Merrell, M., A. Gellhorn, & L. B. Flexner**
   1944. J. Biol. Chem. **153**: 83.
5. **Little, J. M., & J. T. Dameron**
   1943. Am. J. Physiol. **139**: 438.
6. **Cohn, E. J.**
   1945. Science in Progress. **iv**: 319.
7. **Rous, P., H. P. Gilding, & F. Smith**
   1930. J. Exp. Med. **51**: 807.
   Rous, P., & F. Smith
   1931. J. Exp. Med. **53**: 219.
8. **Margaria, R.**
   1930. J. Physiol. **70**: 417.
9. **Landis, E. M.**
   1931. Am. J. Physiol. **98**: 704.
10. **White, J. C., M. E. Field, & C. K. Drinker**
    1933. Am. J. Physiol. **103**: 34.
11. **Keys, A.**
    1937. Trans. Faraday Soc. **33**: 930.
12. **Lewis, T.**
    1927. The Blood Vessels of the Human Skin and their Responses. London
13. **Landis, E. M.**
    1927. Am. J. Physiol. **81**: 124.
14. **Krogh, A.**
    1929. Anatomy and Physiology of the Capillaries. New Haven.
15. **Landis, E. M.**
    1928. Am. J. Physiol. **83**: 528.
16. **Stead, E. A., & J. V. Warren**
    1944. J. Clin. Invest. **23**: 283.
17. **Landis, E. M., L. Jonas, M. Angevine, & W. Erb**
    1932. J. Clin. Invest. **11**: 717.
18. **Danielli, J. F.**
    1940. J. Physiol. **98**: 109.

19. **Zweifach, B. W.**
1940. Am. J. Physiol. **130**: 512.
20. **Rigdon, R. H.**
1940. Surgery **8**: 839.
21. **Landis, E. M., & J. H. Gibbon, Jr.**
1933. J. Clin. Invest. **12**: 105.
22. **Field, M. E., C. K. Drinker, & J. C. White**
1932. J. Exp. Med. **56**: 363.
23. **Hudack, S., & P. D. McMaster**
1932. J. Exp. Med. **55**: 431.
24. **Lewis, T.**
1942. Clin. Sci. **4**: 349.
25. **Lange, K.**
1942. Bull. N. Y. Med. Coll. **5**: 154.
26. **Field, M. E., & C. K. Drinker**
1936. Am. J. Physiol. **116**: 597.
27. **Chambers, R., & B. W. Zweifach**
1940. J. Cell. & Comp. Physiol. **15**: 255.
28. **Brown, E., C. Wise, & J. Wheeler**
To be published.

DISCUSSION OF THE PAPER

Dr. Chambers:

I wish to raise a question, regarding the movement of fluid across the wall of capillary blood vessels. In your 1927 paper (Am J. P. 82), you use the escape of trypan red along the arterial portion of the capillaries to indicate water-escape where the hydrostatic pressure is greatest Rous and his co-workers agreed with you as to identity of dye and water-escape, but differed as to location of the escape, which they claimed to be chiefly at the venous end. Rous and Smith (J Exp Med. 53: 237. 1931.), using lower concentrations of the same dye, trypan red, which is, relatively, highly diffusible, state that, more regularly, dye-escape was greatest at the further end of the capillary, where it should have been least, if pressure changes along the capillary had been the controlling influence. From their experiments, they infer that porosity, rather than pressure differences, accounts for the phenomenon.

In your later publications, you still hold to the Starling hypothesis, and intimate that the observation of dyes escaping from vessels at the venous end of the capillary bed is not an indication of the direction of fluid movement, since an outward diffusion of a dyestuff can occur simultaneously with an inflow of water You would mean, by this, that the dyes used were of sizes below those of the effective blood proteins, so that an osmotic intake of water should occur simultaneously with the outward diffusion of dye along the same region of the blood vessel.

However, Rous and Smith (Ibid. 238) found that the "escape is regularly greatest in the very region (at the venous end) where the concentration (of the dye) in the plasma is least." Surely, this would not be the case, if we accept the assumption that the outward diffusion of the dye is occurring in spite of the fluid inflow! Not only would we have to assume that the outward diffusion of dye is greatest where a fluid inflow is at its height, but that the outward diffusion of the dye is increasing, as its concentration gradient is falling!

On the other hand, if we assume that the porosity of the vessels progressively increases to a maximum at the venous end of the bed, then the greatest escape of the dye at the venous end should indicate the greatest fluid escape. It seems to me that a progressively increasing porosity is also demonstrated by McMaster

and Hudack (J. Exp. Med. 55: 247. 1932.) and McMaster, Hudack, and Rous (J. Exp. Med. 58. 1932.), who found the same gradient, after they had eliminated the effect of hydrostatic pressure by clamping the feeding artery.

Another argument in favor of this seems to me the observation of Smith and Dick (J. Exp. Med. 56: 379. 1932.) that the same gradient of permeability persists, even when an induced hypertonicity of the blood causes an increased absorption of water from the tissues. On the other hand, is it not these hypertonicity experiments on which you base your statement that Rous and his co-workers have shown an outward diffusion of dye simultaneous with inflow of water? It is true that Smith and Dick inferred that the spread of dye from the blood to the tissues cannot be essentially dependent upon fluid movement. They are silent as to the regional disposition of the two movements. However, I think this ambiguity regarding the fluid movement can be explained by the peculiar structure of the capillary bed, in which an outward and an inward diffusion of fluid can occur simultaneously, in the same general region, but *not* in the same vessels. This has been presented in the paper* by Zweifach and myself.

The outward diffusion occurs mainly along the posterior portion of the thoroughfare channels traversing the capillary bed and of the venules; while, under normal conditions, the inward diffusion of fluid occurs into the true capillaries, which are side branches of the thoroughfare channels.

We have used various acid dyes, the diffusibility of which varies with their particle size. On the basis of identifying dye escape with fluid escape, we find that there is fluid escape all along the length of the thoroughfare channels of the capillary bed. The escape progressively increases, until it is most pronounced at the venous end and from the venules.

Dr. Landis:

It seems to me that Dr. Chambers' difficulty arises from his failing to distinguish clearly between three processes: (a) filtration, (b) absorption, and (c) diffusion.

For the capillary network, filtration refers to the mass passage through the capillary wall, from within outward (wherever capillary pressure exceeds the colloid osmotic pressure), of water as a solvent, and of solutes such as sodium chloride, urea, glucose, and even such dyes as are able to pass through the "pores" of the capillary wall. Absorption, as used with reference to the capillary network, refers to the mass passage through the capillary wall, from without inward (wherever the colloid osmotic pressure exceeds the capillary blood pressure), of water as a solvent and of solutes such as sodium chloride, urea, glucose, and even such dyes as may be situated outside the capillary and able to pass through the "pores" of the capillary wall. Filtration and absorption cannot occur, except when different pressures are exerted on a fluid separated by a membrane. Hence, osmotic pressure, capillary blood pressure, and the permeability of the capillary wall are all essential factors in these two processes.

Diffusion, however, refers merely to the movement of a substance, solvent or solute, from a region of high concentration to a region of low concentration, as a result of the random and interfering movement of single molecules. In the absence of a membrane or other barriers, diffusion is "free." In the capillary network, however, diffusion is "impeded," in that the rate of diffusion is slowed, depending upon the relative permeability of the capillary wall. Capillary blood pressure and colloid osmotic pressure should not affect diffusion, except secondarily, where rapid filtration might add to, or rapid absorption subtract from, the effects of pure diffusion.

The studies of Rous indicate that the diffusion of certain dyes is impeded less by the walls of the venous capillaries and venules than by the walls of the true capillaries. A gradient of permeability (to dyes) seems clear, providing vital staining and adsorption have been excluded, as Rous and his co-workers believe to be the case.

Does this mean that simultaneous mass absorption of fluid is excluded in these

* Chambers, R., & B. Zweifach. Ann. N. Y. Acad. Sci. 46 (8): 683–695.

areas? There is an important quantitative difference between the rates of movement produced by the diffusion process and by the filtration-absorption mechanism, when very minute distances of capillary magnitude are being traversed. Thus, Hevesy and Jacobsen (referred to by Landis) have calculated that the diffusion of heavy water through a space of 20 micra is almost complete in 0.1 second. In micro-injection studies, the most rapid rate of absorption observed was $0.06\mu 3/\mu 2$/sec. Assuming that the area of the capillary wall consists of 9/10 stroma and only 1/10 pore space, the maximal rate at which absorbed fluid moved through these pores was only 0.6μ/sec. In other words, it would require about 33 seconds for absorption to move water through a distance of 20 micra, whereas diffusion would accomplish an analogous penetration in 0.1 second. Hence, for the distances involved in exchanges through single capillaries, diffusion may well move molecules of water up to 300 times faster than absorption does. It must be admitted at once that such a great difference may not apply to the large dye molecules as they pass through the capillary wall, but the disparity is still great enough to permit suggesting that absorption of water through the capillary wall cannot impede significantly, for distances of a few micra, the diffusion of a highly diffusible dye in the opposite direction.

The simple observation that a plane is moving rapidly in one direction does not permit the conclusion that there is not even the slightest head wind in the opposite direction. Similarly, the observation that dye is diffusing rapidly through the wall of a venous capillary does not, of itself, exclude the possibility that fluid is being absorbed slowly and simultaneously in the same region. As Peters has mentioned, it is fortunate that the exchange of dissolved substances is accomplished by diffusion, rather than by the much slower process of filtration and absorption.

The implications of these important differences between the filtration-diffusion mechanism and the diffusion process have been considered in detail and need not be repeated at length here. (See **Landis, E. M.** Physiol. Rev. 14: 439–445. 1934; **Peters, J. P.** Yale J. Biol. & Med. 5: 431. 1933.)

It is relevant, moreover, to distinguish clearly between the spotty and early passage of vital red HR, due to filtration from only some capillaries (where capillary pressure is high), and the generalized and slightly delayed diffusion of dye from all venous capillaries which follows slightly later. Filtration of dye-stained fluid from true capillaries occurs abruptly, within the first few minutes after the first dye-stained plasma reaches the capillary network. To quote from the original account (**Landis, E. M.** Am. J. Physiol. 82: 228. 1927): "This relationship holds, however, only during the first ten minutes following the introduction of the dye (vital red HR). After this time the results are obscured by the rapid vital staining of the connective tissue around the smaller venous capillaries and particularly the smallest venules. These are stained deeply, irrespective of pressure and diameter." It was this latter and equally important phase of dye passage which was studied so intensively by Rous and his co-workers.

The dense coloration of the connective tissue, compared to the lighter filtered fluid, suggested selective staining of the venular regions but this interpretation was apparently incorrect, because Rous and his co-workers later provided evidence to exclude selective adsorption. As mentioned in the body of this paper, more evidence is needed before dyes *per se*, can be regarded as completely trustworthy indicators of permeability and of water movement. It is equally true that micro-injection studies of pressure and fluid movement have been restricted to the true capillaries. They should also be extended to include the collecting venous capillaries, venules and thoroughfare channels. The problem of the thoroughfare channel is far from settled and requires further quantitative analysis, with respect to both permeability and capillary blood pressure.

Dr. Chambers:

It is obvious that the concentration gradient determines the direction of diffusion of a diffusible dye-solute. Nevertheless, the flow of the solvent in the opposite direction must impede, at least to some degree, the diffusion of the dye-solute. Conversely, diffusion should become accelerated, when it is in the same

INTERCELLULAR SUBSTANCE IN RELATION TO TISSUE GROWTH

By Eliot R. Clark

Department of Anatomy, University of Pennsylvania, Philadelphia, Pennsylvania

The tissue growth under discussion is that which occurs in adult tissues: such growth as is involved in wound healing, in the late "reparative" stages of inflammation, or in such tissue overgrowths as are found in elephantiasis.

The data have been obtained largely from the direct microscopic observation of growing tissue, as seen in various types of double-walled transparent chambers inserted in the ears, mainly of rabbits,[1, 2] but to a minor extent, of dogs.

What occurs, following the installation of such a chamber in the ear of a rabbit, that involves the two phases of growth, (a) the migration and division phase, and (b) the differentiation phase?

Within a few hours, the Ringer's solution left in the chamber is replaced by an inflammatory exudate.[1, 3] There is usually a complete fibrin network, although this may be absent or incomplete. There are usually many extravasated erythrocytes, single or in masses, but these may be absent. There are always leucocytes, mainly polymorphonuclear (neutrophiles or, in the rabbit, pseudo-eosinophiles), monocytes, and lymphocytes. In the interstices of the fibrin there is clear substance. In some places, this contains suspended cells, which are moved to and fro as a result of changes in the circulation, indicating a fluid consistency; in other places, there is absence of movement, suggesting a viscous or semi-solid condition. This clear material, even when liquid at first, becomes viscous for a distance approximately 3/10 mm. beyond the last circulating capillary, as the new tissue invades the chamber.

That the fibrin may play an important supporting role is indicated in certain chambers in which small growths of epidermis invade the chamber area; for differentiating epidermis, apparently, contains a fibrinolytic ferment, since it is always accompanied by a dissolution of the fibrin,[4] which is followed by a retraction of the tissue. Also, in the dog, fibrin, if present, is lost early, and the new growths of tissue

form slippery, villus-like masses that are moved about freely and tend to remain separate from one another.[5] However, the evidence from our studies is against the view that fibrin is actually transformed into connective tissue fibers. In fact, much of it dissolves when circulating blood capillaries penetrate it.

The visible cellular elements behave, briefly, as follows: Erythrocytes, which, if present, may remain for days out beyond the advancing line of blood vessels, are eventually phagocytized by macrophages. Polymorphonuclears, which accumulate in large numbers, are seen in all stages of active migration, dwindling, and degeneration. After a few days outside the blood vessels, they may die and disintegrate, or they may lose their granulation and dwindle in size, and their nuclei may become spherical; in the latter condition, they may be phagocytized by macrophages, or remain for days as small, inactive, semi-degenerated cells which we have called "dwindled polys."[6] They are not to be confused with lymphocytes, which are very actively migrating cells, although we have a suspicion that such a confusion is very frequent on the part of pathologists. Macrophages, which are the large mononuclears or monocytes of the blood,[7] are always present in large numbers, and phagocytize extravasated erythrocytes and dwindled polymorphonuclears. They move about actively, at first, but after phagocytizing large numbers (probably up to 20) of cells, they become very sluggish and may remain quiescent for weeks, retaining modest phagocytic powers.

This material, apparently, forms an excellent growth-promoting environment, for, into it, there occurs an exceedingly active growth of new tissue derived from the various tissues bordering the chamber space. Usually on the 6th or 7th day, new blood vessel sprouts and fibroblasts may be seen on the edge of the round table, after having bridged a considerable off-table gap, and they grow to the center of the table at a rate which averages between 1/5 mm. and 3/5 mm. per day, until the space over the table, 6.3 mm. in diameter, has a complete layer of newly formed vascularized connective tissue.[3]

While the two tissues mentioned—connective tissue and blood vessels—are consistent invaders of all chambers, other tissues also grow into the space, although less constantly than these two. There are usually varying numbers of lymphatics,[8] still more variable numbers of nerves,[9, 10] both medullated and non-medullated, and there may be ingrowth of epidermis and, less frequently, of cartilage.[11] Also, if bone

has been placed just outside the table area, at the time of the installation of the chamber, there may be migration of bone-forming cells into the chamber and differentiation of bone.[12]

While the tissue is advancing into the uninvaded area toward the center of the table, the second phase of growth, that of differentiation, starts in the older parts of the new tissue and progresses toward the center. This has been nicely worked out by M. L. Stearns, in a study of the growth and differentiation of connective tissue in transparent chambers[13] She found that differentiated connective tissue fibers became visible in connection with cells, in tissue which had formed 4 to 6 days earlier. We have observed nerve-supplied, smooth muscle cells contracting on arterioles which were 9 days old.[9] In this older part of the new tissue, where differentiation is occurring, there may be no further new growth of the migration and extension phase.

The new tissue, as it grows and differentiates, gradually displaces the original material. The fibrin is either dissolved or phagocytized by the macrophages, which digest it; much of the fluid or viscous extracellular material is replaced.

It seems obvious that the original material, which consists of an inflammatory exudate, contains growth-promoting properties, physical and chemical, which promote the migration and division phase of growth; that these properties are lost after the new tissue has formed; and that there are then present properties which favor the differentiation phase of growth.

This diminution of growth-promoting properties has, at times, manifested itself before the table area of the chamber has been completely occupied by new tissue, while, on the other hand, growth-promoting properties have been reintroduced by a renewed period of mild inflammation, as indicated by the following:

In the growth of tissue from the periphery to the center of the round table chamber, there are differences in the rate and amount of growth, if the tissue in the chamber is undisturbed. Over a period of years, the construction of chambers has changed in the direction of making them more and more rigid, and better protected against trauma. Thus, there has been an unintentional experiment on the presence and absence of repeated mild inflammatory reactions. In earlier chambers, there were repeated traumatizations, each one resulting in mild injuries and mild inflammatory reactions. The introduction of the shields and splints very greatly reduced the trauma. The stiffening of the chambers by doubling the celluloid prevented warping and made the cham-

ber still more stable. We finally arrived at chambers in which, occasionally, the entire growth phase was completed without any renewal of the original inflammatory exudate. A comparison of the rates of growth in these various chambers indicates that, in the earlier types of chambers, in which there were repeated injuries with ensuing mild inflammations, active growth continued without appreciable diminution in rate, until the entire space was filled, and continued in the interstices and over the top, as the inflammatory edema raised the cover of the chamber, increasing the depth of the space. On the other hand, in the later chambers in which successive inflammatory reactions were largely eliminated, while growth started at the same rate as in the earlier chambers, there was a noticeable reduction in the rate of growth before the vascularization of the table area was complete. There was no new growth, thereafter, unless and until a new mild inflammatory exudate was produced.

In one such chamber, studied intensively, the rate of growth diminished from day to day, until the last vessels to form in the center of the round table took several days to cover a distance which would have been covered, earlier, in one day. Furthermore, while the depth of the space measured approximately 100 micra, the growth was restricted to a bottom layer which filled only about 2/5, or 40 micra, of the depth of the space, leaving a layer of clear, uninvaded material of a fluid or viscous nature intervening between the tissue and the top of the chamber. This tissue and the uninvaded space above it persisted without change for 40 days, during which time there was no inflammatory reaction. Then, a slight infection nearby caused a mild inflammation which continued for 3 or 4 days. There was increased rate of blood flow, some sticking and emigrating of leucocytes, movement of cells in the material above the tissue showing a definite fluid condition: all evidences of inflammation and the formation of a mild inflammatory exudate. On the third day, there were new blood vessel sprouts and young fibroblasts growing in a layer superficial to the tissue. This growth continued for about 3 days, at which time a new layer of tissue had been formed, consisting of new circulating vessels and fibroblasts. In the meantime, the inflammatory condition subsided, growth subsided, and a new layer of tissue had been added, but there still remained an uninvaded space of a depth of about 40 micra. Several weeks later, there had been no noticeable change, when there occurred a second mild inflammation, caused by several days of unseasonably warm weather. The growth phase was repeated, and this time the new tissue filled the remaining space.

Thus, a mild inflammatory exudate provides a growth-promoting environment, which brings about the migration and division or proliferation phase of growth. If the inflammation is of brief duration, the growth is slight; if extensive and prolonged, as at the installation of the chamber, the growth is prolonged and extensive. The characteristics which make this a growth-promoting medium are, apparently, gradually lost. Whether specific substances are there, which are gradually appropriated or neutralized by the growing tissue, or whether there is some other explanation, it is impossible at present to decide. There is every indication and some direct evidence that, in the quiescent growth phase, the uninvaded, clear material is high in protein, and yet without any growth-stimulating potency.

What part, if any, do lymphatics play in this new growth of tissue? There is great variation in numbers of lymphatics and in the time at which they appear in our chambers.[8] Sometimes, they grow along with the most advanced tissue; in other cases, not until much later. There is no appreciable difference in the rate of advance of new tissue, whether lymphatics are present during the growth—even when they have open tips—or whether lymphatics fail to appear in the field of growth until after the space has been completely occupied by new tissue.

One very conclusive observation has been described,[*] in which the lymphatics had, for several days, tips open to the tissue space. There was fluid, containing suspended erythrocytes and leucocytes, which moved freely into the lymphatic, sometimes well along it, off the table area of the chamber. There was also back flow—the fluid and suspended cells rushing out again. In short, there was an apparent perfect arrangement for draining the tissue space and thus preventing new growth of tissue. Nevertheless, growth continued undiminished.

In commenting on these observations, Drinker and Yoffey (p. 71[14]) mistakenly describe our chambers as without any forces which might cause movement of lymph. This is not the case, for, as we stated: "Individual erythrocytes, bobbing freely in the fluid of this uninvaded area, were seen to enter the lymphatic, where they moved along for a relatively long distance, occasionally even passing off the table, into the communicating vessels of the preformed tissue. . . . The 'bobbing' of cells was caused by the heart beat, while the extensive movement of cells was the result of dilatation and contraction of arteries which produced a raising, followed by a lowering, of the mica cover. When

[*] Clark, E. R., & E. L. Clark. Am. J. Anat. 52: 285. 1933.

raised, the cells moved out the peripheral ends of the lymphatics, and when lowered they were forced from the extravascular space into the lymphatics."

Since there are periodic contractions and dilatations of the arteries and arterioles of the rabbit's ear, occurring on the average twice each minute, and since each alternation produces an effect upon the fluid in the lymphatics in the chambers, it would seem that in few locations in the body are the lymphatics subjected to a greater amount of squeezing and massage.

Lymphatics, in general, play a role that seems negligible in relation both to growth and to normal tissue activity. There is no noticeable difference in the appearance or behavior of a tissue, whether lymphatics are present or not, or whether they are present and functioning, or present and not functioning. Both formed elements and unformed fluid or viscous material disappear at the same rate, regardless of whether or not lymphatics are present. Indeed, debris such as dwindled polymorphonuclears and dilapidated erythrocytes are disposed of by macrophages much more rapidly and easily in the tissue spaces than in the peripheral lymphatics. We have frequently seen accumulations of erythrocytes and dwindled polymorphonuclears remain in lymphatics for many days. If the lymphatics are pressed upon so that the walls are broken and these cells forced out into the tissue spaces, the cells are promptly (usually within 24 hours) picked up and disposed of by macrophages. It is true that macrophages may enter the lymphatics and phagocytize degenerated cells there, but it is rare that, following hemorrhage into the lymphatics or massive immigration of polymorphonuclears during an inflammation, enough macrophages enter the lymphatics to take care of more than a very small proportion of the debris. The endothelial cells, themselves, of peripheral lymphatics are not phagocytic, nor do they give rise to phagocytic cells. The latter come from the large mononuclears of the blood stream.

What part might lymphatics play in excessive tissue growth, such as occurs in elephantiasis? If lymphatics were blocked on a large scale, so that debris accumulated within them, as we have seen on a small scale, it would seem that they might furnish a sort of culture medium, held within a miniature test tube formed by the lymphatic itself, which could favor the growth and survival of bacteria, thus increasing the formation of inflammatory exudates that provide a growth-environment. There does not seem to be proof, as yet, that overgrowth of tissue re-

sults solely from the accumulation of excess protein in the tissue spaces which, by itself, provides a growth-promoting environment. That view has been strongly suggested by Drinker and Field[15] and by Drinker and Yoffey,[14] although in both books it is admitted that overgrowth of tissue may be accelerated by the presence of local infection.

As to the nature of the intercellular substance, much, clearly, remains to be learned. There is evidence, from studies of the tadpole's tail and from observations on the ear chambers, that, under normal conditions in these two regions, the intercellular substance is of a viscous nature, rather than in the form of free fluid. This view is supported by the studies of Sylvia H. Bensley,[16] who found the material between connective tissue fibers to be viscous, with staining properties that indicated a partial mucous content. Suggestive, in this connection, is the restricted spreading of substances such as India ink injected into the skin; the increased spreading that occurs in the Duran-Reynals reaction,[17] when testicular extract is injected simultaneously with India ink; and the finding of Chain and Duthie,[18] that testicular extract contains a mucolytic enzyme.

While the intercellular substance may be of a mucous or gelatinous nature, normally, its fluid content can be increased in a very short time, from the action of factors which either increase markedly the amount of fluid diffusing out from the blood capillaries, or which interfere with the return of fluid from the tissues.

R. G. Abell,[19] using the "moat" chamber, has studied the rate of diffusion of nitrogenous substances through the walls of capillaries and their accumulation in a small outside receptacle called the *moat*. He has found a definite passage of proteins, which is greater in the case of newly formed than of older capillaries. If the material in the moat is not disturbed, it becomes more and more viscous, until it may reach a semi-solid state. This appears to simulate the formation of a viscous intercellular substance. No studies of this substance have been made, aside from the analyses for protein and non-protein nitrogen, and they show a gradual increase in protein nitrogen. Further analysis may help to clear up some of the questions concerning the nature and formation of intercellular substance.

Regarding the possible presence of specific growth-promoting substances in the mild inflammatory exudate, we have no data to contribute. Tissue culture studies suggest the possibility of albumoses—perhaps mixed with blood proteins. Albumoses could conceivably diffuse as such through the endothelial wall, or form outside the vessels

from blood or tissue proteins, by the digestive action of enzymes. While the future will undoubtedly yield the complete picture, the knowledge that the material formed during a mild inflammation constitutes a growth-promoting medium provides the surgeon and physician with information of therapeutic importance.

REFERENCES

1. **Sandison, J. C.**
   1928. Am. J. Anat. **41**: 447.
2. **Clark, E. R., H. T. Kirby-Smith, R. O. Rex, & R. G. Williams**
   1930. Anat. Rec. **47**: 187.
3. **Clark, E. R., W. J. Hitschler, H. T. Kirby-Smith, R. O. Rex, & J. H. Smith**
   1931. Anat. Rec. **50**: 129.
4. **Clark, E. R., & E. L. Clark**
   1944. Anat. Rec. **88**: 426.
5. **Moore, R. L.**
   1936. Anat. Rec. **64**: 387.
6. **Clark, E. R., E. L. Clark, & R. O. Rex**
   1936. Am. J. Anat. **59**: 1.
7. **Clark, E. R., & E. L. Clark**
   1932. Am. J. Anat. **51**: 49.
8. **Clark, E. R., & E. L. Clark**
   1932. Am. J. Anat. **51**: 49.
   1933. Am. J. Anat. **52**: 273, 285.
   1937. Am. J. Anat. **60**: 253; **62**: 59.
9. **Clark, E. R., E. L. Clark, & R. G. Williams**
   1934. Am. J. Anat. **55**: 47.
10. **Clark, E. R., & E. L. Clark**
    1938. Anat. Rec. **70**: 14 (Suppl.).
11. **Clark, E. R., & E. L. Clark**
    1942. Am. J. Anat. **70**: 167.
12. **Sandison, J. C.**
    1928. Anat. Rec. **40**: 41.
    Kirby-Smith, H. T.
    1933. Am. J. Anat. **53**: 377.
13. **Stearns, M. L.**
    1940. Am. J. Anat. **66**: 133; **67**: 55.
14. **Drinker, C. K., & J. M. Yoffey**
    1941. Lymphatics, Lymph and Lymphoid Tissue. Cambridge.
15. **Drinker, C. K., & M. E. Field**
    1933. Lymph and Tissue Fluid. Baltimore.
16. **Bensley, S. H.**
    1934. Anat. Rec. **60**: 93.
17. **Duran-Reynals, F.**
    1928. C. r. Soc. de Biol. **99**: 6.
    1929. J. Exp. Med. **50**: 327.
18. **Chain, E., & E. S. Duthie**
    1939. Nature **144**: 977.

19. **Abell, R. G.**
1939. Collecting Net **14**: (10).
1940. Anat. Rec. **76** (Suppl. 2.): 1.

DISCUSSION OF THE PAPER

Dr. Drinker (*School of Public Health, Harvard University, Boston, Mass.*):

Elephantiasis occurs in the dog when the lymphatics are wiped out and the leg is dependent. For this to develop, one needs a dependent part, an active animal, and some degree of inflammation. It would seem that those conditions are not present, in the rabbit ear chamber. It is not surprising that there is little movement of lymph in such chambers. The only force for such movement would seem to be coming from other vessels.

Dr. Valy Menkin (*Department of Pathology, Duke University School of Medicine, Durham, N. Carolina*):

It would appear from our experiments that injured cells release a growth-promoting substance in inflammation. Such growth stimulation is not produced by injections of blood into the tissues, but, in our studies, it did occur, following injection of inflammatory exudate.

Dr. Richard Abell (*University of Pennsylvania, Philadelphia, Pa.*):

In connection with the formation of intervascular or interstitial substance, mentioned by Dr. Clark, it may be of some interest to note that, in transparent chambers which contain a reservoir that communicates freely with the region of growing tissue, there is a continued increase in protein content and viscosity of the material in the reservoir. Such reservoirs are, at first, filled with Ringer-Locke solution, and the increase in viscosity is due to accumulation of substances which continually diffuse into the reservoirs from the growing vessels, or at least principally from the vessels. One cannot help wondering whether a similar increase in viscosity may not have occurred outside of the growing vessels in the chambers described by Dr. Clark, and whether such changes in character of the material surrounding the vessels, if they occurred, might not have been related to changes in growth rate or pattern.

In studies carried out under Dr. Clark's direction (Abell, 1942), it was shown that blood capillary sprouts allow substances to pass more freely through their walls than do mature capillaries, and this may also have some bearing upon the character of the newly formed intervascular substance of growing tissue in the transparent chambers.

One more point may be of some interest: namely, that, when testicular extract, containing the Duran-Reynals "spreading factor," is introduced into transparent chambers and allowed to diffuse to the region of the vessels, the capillaries become more permeable to the blue dye T.1824, intravenously injected (Abell & Aylward, 1941). Since such extracts contain hyaluronidase, and since there is much evidence to show that this hyaluronidase is responsible for the "spreading reaction" in the subcutaneous tissue produced by such extracts, it seems not unlikely that it is also responsible for the increase in capillary permeability. Thus, hyaluronidase, which has been shown to hydrolyze the hyaluronic acid of intervascular substances, also renders the capillaries more permeable and allows more substances to pass through their walls into the surrounding tissue. The point of particular interest here would seem to be that a substance (hyaluronidase) which renders the intervascular material more permeable (spreading reaction), also renders the blood capillaries more permeable. Similar findings were reported by Duran-Reynals, in 1939.

Dr. Chambers:

It would be a good thing to differentiate between the two terms, "intercellular cement" and "intercellular (interstitial) substance." A decrease in the calcium-

content of the perfusing fluid within the capillaries causes a deficiency in the intercellular cement, but no change in the intercellular substance. On the other hand, a deficiency of vitamin C causes a deficiency in the interstitial substance, but not in the intercellular cement.

Dr Warren H. Lewis (*The Wistar Institute, Philadelphia, Pa*)

In tissue cultures, macrophages drink in (pinocytose) many times their own volume of fluid in relatively short periods of time The globules of fluid thus taken in contain stuff that cannot diffuse into cells After digestion, the altered fluid diffuses out of the cell into the surrounding medium The same process occurs *in vivo* and, since macrophages occur in great numbers throughout the body, it probably plays an important role in determining the character of the interstitial substance (ground substance of the connective tissues)

Dr Stuart Mudd (*University of Pennsylvania, Philadelphia, Pa*)

The pneumococcus capsule contains hyaluronic acid which gives it its open jelly-like structure I wonder whether the intercellular substance does not have the same property for the same reason?

CONDITIONS IN THE SKIN INFLUENCING INTERSTITIAL FLUID MOVEMENT, LYMPH FORMATION, AND LYMPH FLOW

By Philip D. McMaster

Laboratories of The Rockefeller Institute for Medical Research, New York, N Y

Part of the fluid which escapes from the blood vessels into the tissues returns to the blood directly, part enters the lymphatic capillaries to become lymph. Little is known about the formation and flow of lymph, that is to say, of the factors influencing the movements of interstitial fluid through the tissues into the lymphatics and of the conditions favoring or hindering the flow of fluid in the vessels themselves.

To throw some light on these subjects, studies were made of the movements of fluid introduced into cutaneous connective tissue, and of the pressure relationships within the lymphatic capillaries and in the tissues outside these vessels. Techniques have also been devised to determine changes in lymph flow in living human and animal skin and to observe some of the activities of the lymphatics, in the repair of injuries and in the defense of the body against infection. For the present paper, we have selected from these studies the findings which are most closely related to the problems under discussion in this conference. Much of the work has already been published; the remainder stands unfinished and incomplete, owing to sudden interruption, more than five years ago, by the approach of the recent world catastrophe.

THE INTERMITTENT TAKE-UP OF FLUID FROM CUTANEOUS TISSUE

To learn more about the formation of lymph, almost microscopic amounts of certain fluids were introduced into the connective tissues of skin, and their absorption at atmospheric pressure and movement through the tissues under various low pressures were directly measured in a variety of physiological and pathological states. At the outset, it was clear that studies of the sort could not be carried out, if the fluids introduced into the tissues were allowed to enter directly into torn blood vessels or lymphatics. Nor would the pressures exerted upon the fluids

to bring about their movement through the skin yield evidence of the true pressure conditions within the tissues, if the fluids were employed in easily measured amounts, sufficient to distort the formed elements and create artificial interstitial pressures. Consequently, means were found to avoid these difficulties. By the introduction of fluid into the cutaneous tissue in such a manner that it did not usually enter directly into the vessels, and in such minute amounts that its movement through the tissues approximated the natural, artificial intercellular pressures were either reduced to a minimum or created at will.

Methods and Techniques

The method employed to measure the very minute amounts of fluid brought into contact with the dermal tissues of small animals, and taken up by the skin, has been fully described and illustrated, elsewhere.[1] Here, only certain important principles of the techniques need be outlined. Briefly, the apparatus consisted of a 0.2 cc. pipette, connected by several three-way stopcocks with a gauge 30 platinum-iridium needle, at one end, and at the other, with manometers of varying sensitiveness and a by-pass to room air, so that the test fluids could be brought into the tissues at atmospheric pressure and later subjected to known low pressures. Since the take-up of fluid, by the skin, from this smallest of needles was slight indeed, it was necessary to submerge the injecting apparatus in a constant temperature bath and to observe the movements of the fluid in the pipette through a microscope, itself partly submerged.

The studies were carried out on the skin of the ear, backs, flanks, and legs of mice, guinea pigs, and small rabbits. The ear of the mouse was found most satisfactory for our purposes, since the smallest vessels are visible in the organ and much vascular injury can be avoided by placing the injecting needle in tissues midway between the large radiating blood vessels and parallel to the course of the lymphatics.

To bring the test fluid into contact with the connective tissue of living skin, in such a way that it usually failed to enter blood or lymphatic vessels directly, the following techniques were used. With a dissecting needle ground to less than 0.2 mm. in diameter, and under a binocular microscope, a tunnel parallel to the surface of the skin and 3 to 4 mm. in length was made through the tissues of the subpapillary layer of the corium. Thus prepared, the animals were transferred to a glass dish standing in the water of the constant temperature bath and equipped with portholes sealed by rubber tissue through which the in-

jecting needle penetrated to reach the skin. The needle, the pipette, and the skin area to be injected, all lay in the same horizontal plane. Under a second binocular microscope, the needle was slowly pushed into the tunnel in the skin to reach its further end, and finally adjusted in place, so that there were no visible signs of tension. Since the shaft of the injecting needle had a much larger diameter than the dissecting needle used originally to open the pathway through the tissues, blood or lymphatic capillaries which might have been torn in making the tunnel were occluded.

Next, a hair-like gauge 40 steel wire, ground to a rounded tip, was inserted into a rubber tube filled with the test fluid and connected to the needle. Under the microscope and with the utmost care, the blunt tip of the wire was used to push the tissues away from the needle's tip, thus forming a minute cavity into which the test fluid entered at atmospheric pressure. Finally, by appropriate manipulation of the various stopcocks, the test fluid within the tissues and in the needle was brought into communication with that in the measuring pipette, also at atmospheric pressure. An observer seated at the binocular microscope watched continuously for movement of the meniscus in the pipette, with respect to its position in relation to two ocular micrometer scales situated in the eyepieces. The amount of movement, if any, was recorded at ½ minute or 1 minute intervals.

The success of the work depended wholly upon the proper placement of the needle in such a way that no fluid could pass directly to or from torn lymphatics or blood vessels. Accordingly, special tests were employed in each experiment to determine whether or not this had been accomplished. If blood cells, indicating hemorrhage, appeared in the tissues, or if any signs of tissue injury or distortion of tissue elements became visible under the microscope, the experiment was abandoned at the outset. In scores of tests, solutions sufficiently colored to be visible in blood capillaries and lymphatics were employed as test fluids. In more than a hundred experiments, these fluids were seen to enter directly into lymphatics in less than 4 per cent of the trials, and appeared in the blood vessels in less than 1 per cent.

At the end of each experiment, while the needle remained in the skin, an aqueous isotonic solution of a blue vital dye, pontamine sky blue, was injected intravenously into the animal. If the dye did not appear as promptly in the smallest blood vessels and capillaries at the very edge of the cavity at the needle's tip as in capillaries elsewhere in the ear; or if dye-colored blood could not be seen flowing through capilla-

ries that lay between the level of the needle and the surface of the skin; or if the rate of blood flow close to the needle, as judged by the speed of coloration of the vessels there, was not like that in other parts of the ear, the experiment was discarded. Ecchymoses of dye, such as would have formed, had any of the blood vessels or capillaries torn by the needle remained open, were seldom seen, and, on the occasions when they did appear, these experiments too were discarded. Further, if dye escape from blood vessels into the tissues surrounding the needle was more intense than in other parts of the ear, the findings in that experiment were not considered significant, since the phenomenon has been shown by work from this laboratory[9, 10, 15-17] to indicate injury of blood vessels.

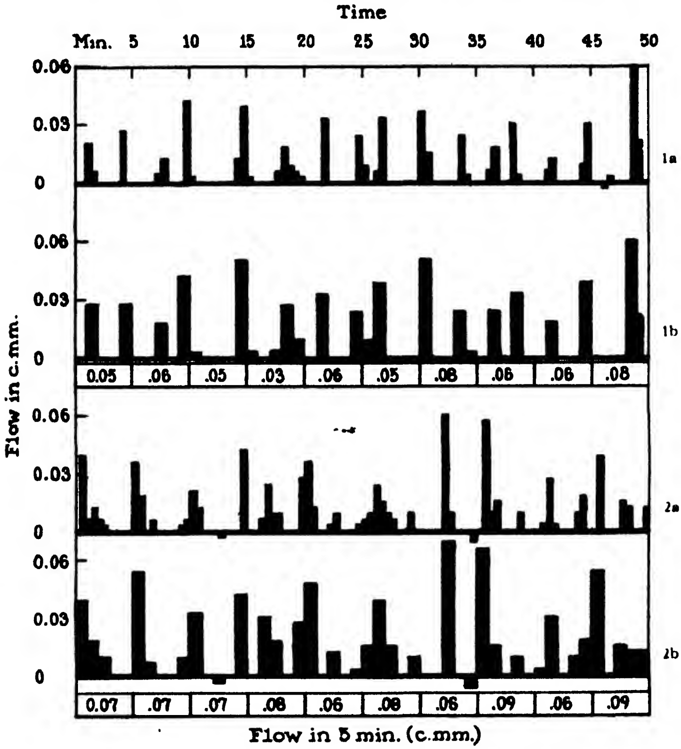
The Movement of Fluid into Cutaneous Tissues at Atmospheric Pressure

In more than four hundred tests, made in the manner just described, either Locke's or Tyrode's solution, when brought into contact with the connective tissue of skin at atmospheric pressure, was taken up by the tissues very slowly, but with a surprising intermittency. Short periods of inflow into the skin, some lasting but 10–15 seconds, some for approximately a minute, and a few for several minutes, were rapidly succeeded by periods of no inflow. Flow usually ceased and began abruptly. As will appear below, there was no such intermittency in control experiments which tested the character of equally slow flow through the apparatus alone.

In FIGURES 1 and 2, the intermittent inflow of fluid into the skin has been charted for two typical experiments. FIGURES 1a and 2a show the amounts of inflow which took place during each ½ minute interval, and in FIGURES 1b and 2b the flow is recorded at 1 minute intervals. The latter figures have been included to furnish a comparison with data presented below and obtained from experiments in which we were unable to make readings at intervals shorter than 1 minute. In the figures, each of the solid black columns standing above the base line represents the amount of inflow during that particular ½ minute or 1 minute period. Columns extending below the line depict backflow.

As already stated, an observer watched continuously the movement of the meniscus in the pipette and recorded the amount of movement observed at the end of each 30 second or 1 minute period. It is to be stressed, therefore, that the columns in the figures represent merely these recordings, that is to say, the amount of fluid that had entered

the tissues at the end of a 30 second or 1 minute period. The columns give only the roughest indication of the curve of the changing rates of flow and of precisely when it began and stopped. Nevertheless, on



FIGURES 1 and 2. The intermittent entrance of Locke's solution, at atmosphere pressure, into living skin. The readings have been plotted at 30 second intervals (FIGURES 1a and 2a) and at 1 minute intervals (FIGURES 1b and 2b).

many occasions, the movement of fluid was sufficiently great to show that the beginning of flow and its cessation were abrupt. The matter has been fully discussed in a previous paper.[1]

The take-up of fluid during each 5 minute period is given for each experiment in all of the figures. In most, the rate of inflow at atmospheric pressure was surprisingly constant, varying in more than 80 per cent between 0.04 and 0.08 c.mm. per 5 minutes. In a few instances as much as 0.1 to 0.12 c.mm. entered the tissues in 5 minutes, and in very few there was no flow.

Occasionally, intermittent inflow did not appear. Instead, in about 3 per cent of the trials, a continuous, irregular inflow occurred at about the same rate as in the other animal experiments, but without evidence of periodicity. The irregularities which appeared were like those observed in the controls, to be discussed below. In another 3 per cent of our experiments, there was no flow at atmospheric pressure. In such instances, the pressure of a column of water, 1 or 2 cm. in height, applied to the contents of the pipette, usually initiated flow. If none occurred, the needle was tested for obstruction as described elsewhere.[1] All instances of obstruction were ruled out.

In the experiments so far considered, no edema of the skin was visible under the microscope. In many other instances, some minutes after placing the needle in the skin, edema appeared. In these, the intermittent inflow of fluid halted and intermittent backward movement began from the tissues into the apparatus.

The Movement of Fluid into Cutaneous Tissues under Low Pressures

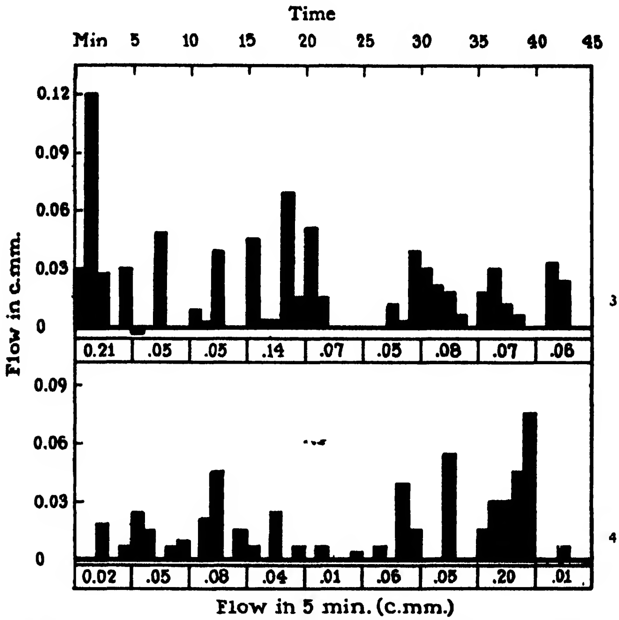
Locke's solution, brought into contact with the connective tissue of normal skin and then forced into it under pressure, continued to enter it in a periodic manner at all pressures less than 4 to 4.5 cm. of water. Figures 3 and 4 show that periods of no movement and abrupt flow into the skin of the mouse's ear still made their appearance, when pressures of 2 and 1 cm. of water, respectively, were brought to bear on the fluid introduced by the needle.

Control Experiments

The findings suggested that the intermittency of flow might be produced by physical forces within the apparatus, rather than by physiological happenings within the tissues. Accordingly, an elaborate series of control experiments were made to determine the point. These have been fully described before,[1] and only the briefest indications of their principles will be given here.

In the animal experiments, the Locke's solution was drawn through the pipette and needle by forces of an unknown nature, while the needle's tip lay in a pool of fluid within the tissues. To study the characteristics of flow through the apparatus alone, at the same rate at which it occurred in the animal experiments, or at slower or faster rates, the needle was placed in a pool of Locke's solution and the fluid drawn through the apparatus, as it was in the animal experiments, by changing the osmotic conditions at the needle's tip. In other tests, the

fluid was forced through the pipette by very slight pressure. Figures 5a and 5b to 8a and 8b indicate the characteristics of the flow through the apparatus alone, at the same rate as in the animal experiments, or



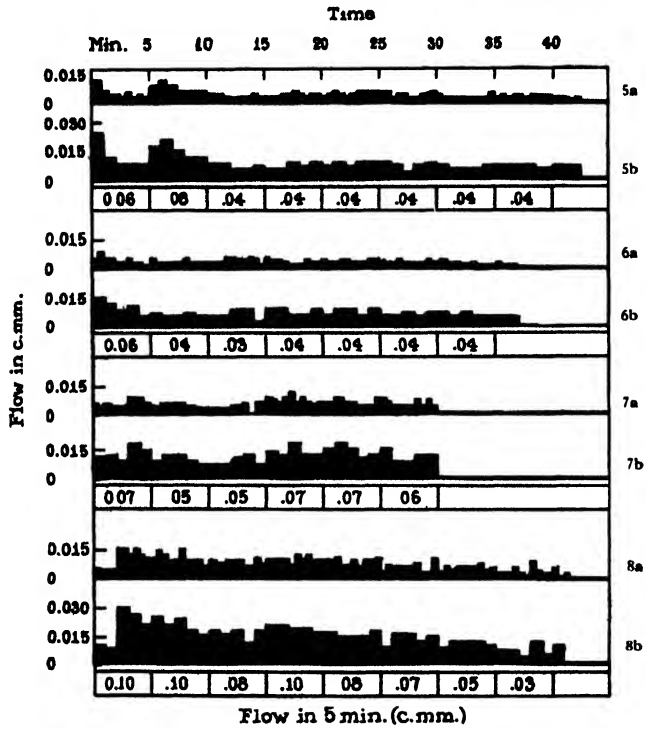
Figures 3 and 4. The intermittent interstitial movement of Locke's solution forced into living skin by pressures of 2.0 and 1.0 cm. of water, respectively. Readings from two different experiments are shown, at minute intervals only.

at faster or slower rates. For each test, the flow has been charted at half-minute and at minute periods, and, as in the preceding figures, the total flow occurring in each 5 minute period is given below each drawing. In these control tests, the flow showed marked irregularities, probably caused by physical forces in the apparatus, but it was continuous and the characteristic intermittent movement seen in the animal experiments was lacking.

FACTORS INFLUENCING THE INTERMITTENT PASSAGE OF LOCKE'S SOLUTION INTO LIVING SKIN

Since the intermittent flow through the apparatus occurred only in the animal experiments, and not in the controls, the phenomenon was not to be explained by physical forces active in the pipette. Some phy-

siological effect seemed to be responsible. Among the explanations for the phenomenon which suggested themselves, two seemed worthy of investigation: it might be ascribed either to recurring alterations



FIGURES 5-8 record the flow of Locke's solution through the pipette and needle of the injecting device, during control experiments in which the osmotic forces were used, as described in the text, to draw fluid through the injecting device. FIGURES 5a-8a show the movement of the meniscus plotted at ½ minute intervals. In FIGURES 5b-8b, we have plotted the movement at 1 minute intervals. The slight irregularities as plotted contrast strongly with those shown in FIGURES 1-4, which were obtained in animal experiments.

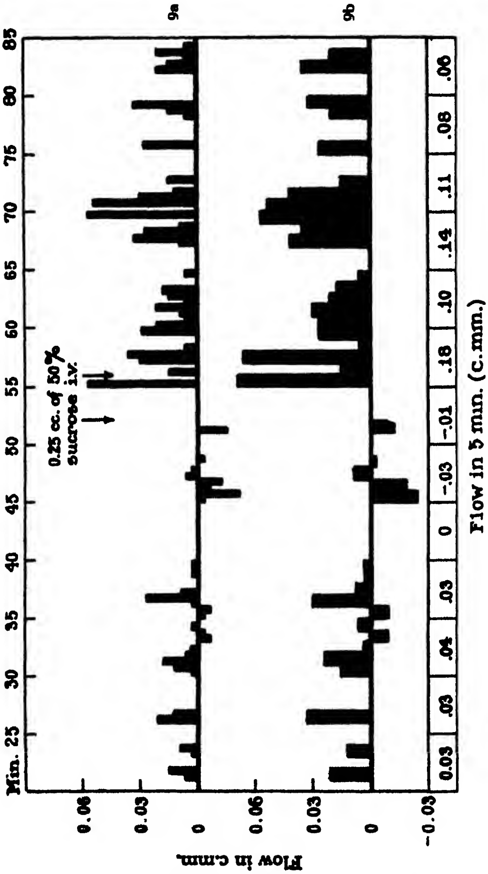
in the vascular or tissue conditions determining absorption; or there might have been an intermittent expansion and contraction of tissue elements, which allowed fluids to move through the tissues in an irregular manner. If fluid is absorbed into the blood intermittently under natural conditions, the present concepts of the mechanics of fluid exchange require modification. If periodic changes take place in the tissues, affecting the movements of extravascular fluid, then it becomes a matter of interest to know what these changes are.

Before attempting to learn which of these two possibilities might account for the intermittent character of the entrance of fluid into the skin, it was clearly necessary to determine whether or not the methods employed were sensitive enough to demonstrate changes in the rate of absorption of fluid introduced interstitially. To test the point, a variety of experiments was made in which Locke's solution, at atmospheric pressure, was brought into contact with the dermal tissues of the ears of normal mice, and after the rate and characteristics of its inflow had been observed, 0.2 or 0.3 cc. of 50 per cent sucrose solution, which increased the osmotic pressue of the blood, was injected into a tail vein, while the Locke's solution continued to flow into the ear. The injection of sucrose into a tail vein greatly increased the take-up of Locke's solution by the connective tissue of the ear.[2] In further experiments, it was shown[2] that pain and tactile stimuli, either from the injecting needle or from manipulation of the tail, did not account for the characteristic intermittency of the take-up fluid by the skin. Clearly, the methods employed detected the abnormal absorption of small quantities of interstitial fluid from the skin. In these experiments, in which the volume of fluid circulating in the blood was temporarily increased, the circulation in the ear seemed more complete than before and the vessels were dilated; nevertheless, there was an increased entrance of fluid into the skin. The findings enabled one to suppose that the entrance of Locke's solution into the skin, under the ordinary circumstances of experimentation, was the result of an intermittent absorption of fluid by the blood, although a series of tissue changes admitting fluid into the tissues irregularly could not be absolutely ruled out.

Further experiments supported this explanation of the findings. As already mentioned and described in preceding papers,[1, 2] in about one-fourth of our experiments the insertion of the injecting needle into the ear of the mouse led, after some minutes, to the development of edema of the skin. In all experiments in which the ears became edematous, the intermittent entrance of fluid at atmospheric pressure ceased, and sooner or later backflow into the injecting device occurred. The backflow was intermittent in character, like the inflow. In these instances, intravenous injections of sucrose led to a renewed inflow of fluid from the injecting apparatus, showing that the extravascular fluid had been absorbed and that, in addition, greater amounts of Locke's solution had been taken up from the apparatus.

Figures 9a and 9b give the data obtained between the 20th and 85th

minutes, during an experiment of this kind, as recorded at intervals of 30 seconds and one minute, respectively. For 30 minutes, Locke's solution, brought into contact with the dermal tissues of the ear at



FIGURES 9a and 9b. An intravenous injection of hypertonic sucrose solution reverses the flow of fluid as the ear becomes edematous. As explained in the text, Locke's solution, which had been flowing into the skin at atmospheric pressure in an intermittent manner, began to move backwards intermittently into the needle as edema of the ear developed. An intravenous injection of 50 per cent sucrose solution caused fluid again to enter the skin. The findings are shown in FIGURE 9a as read at ½ minute intervals, and in FIGURE 9b as read at 1 minute intervals.

atmospheric pressure, entered the skin in the usual manner. Between the 30th and 40th minutes, the superficial skin near the needle became obviously edematous. From the 33rd to the 40th minute, alternate backflow and inflow of fluid occurred through the needle. From the 40th to the 46th minute, no entrance of fluid took place, and then, in the next 7 minutes, two periods of backflow occurred, as the chart shows. During the 53rd minute, an intravenous injection of sucrose was started

and 0.25 cc. given in 4 minutes. An irregularly intermittent inflow began before the injection was completed, and continued for 20 minutes. Thereafter, the entrance of fluid into the skin continued at a much faster rate than at the beginning of the experiment, before the injection of sucrose had been given and before the edema had made its appearance.

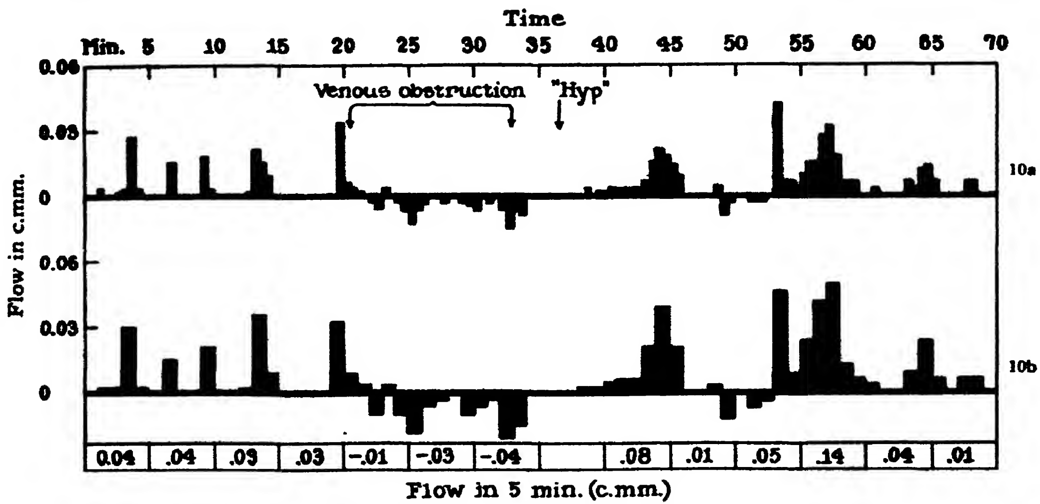
The Effect of Hyperemia

To test whether or not intermittent expansion or dilatation of tissue elements accounted for the intermittent take-up of the fluid rendered available to the tissues by our apparatus, moderate and extreme dilatation of the blood vessels of the ears was induced by a variety of methods. Fluid, at atmospheric pressure, in contact with a tissue which is becoming hyperemic, should be pressed upon by the dilating vessels and forced back into the pipette, if the intermittent entrance seen under normal circumstances is due to changes in the tissues. As the tests showed, the development of hyperemia in the ear was accompanied by greater inflow, despite the fact that both larger and smaller blood vessels became dilated.

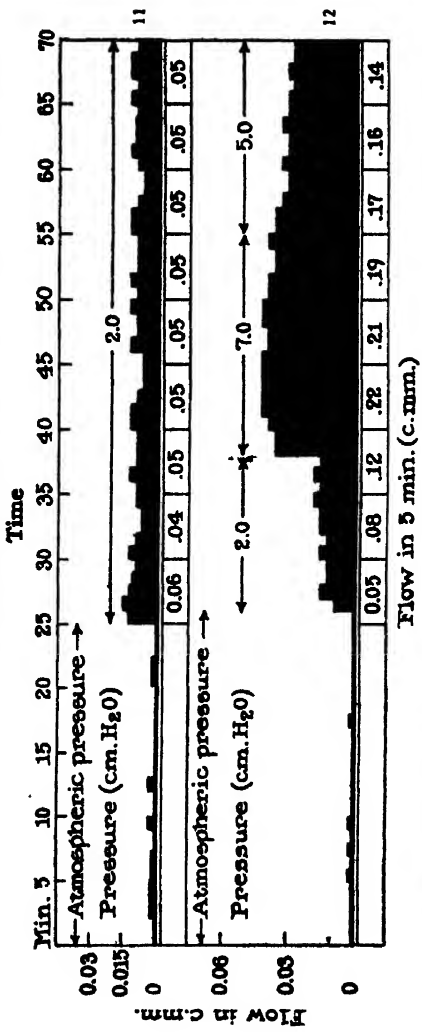
The Effects of Venous Obstruction and of Subsequent Reactive Hyperemia

The release of temporary venous obstruction to a part of the body is followed by intense reactive hyperemia,[13, 18, 19] and, as work from this laboratory has shown, lymph flow is greatly increased.[13]

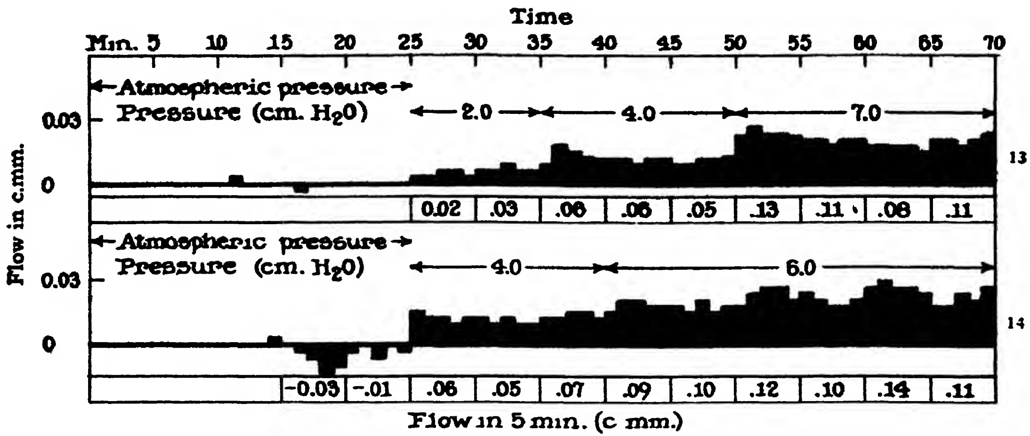
In six experiments, Locke's solution at atmospheric pressure was introduced into the skin of rabbits' ears, which were subjected, about 20 minutes later, to temporary venous obstruction. Figures 10a and 10b give typical findings, plotted at intervals of 30 seconds and one minute. Venous obstruction stopped the usual intermittent intake of fluid at atmospheric pressure, and free fluid collected in the tissues and passed backwards into the pipette, as indicated by the columns extending below the base level. Both the inflow and the backflow were intermittent, from which it follows that neither the intake nor output of fluid in localized regions is a continuous process. Upon the release of venous obstruction, an intense reactive hyperemia with great dilatation of the vessels, appeared, as shown in the figure by the arrow labeled "Hyp." The dilatation of the vessels did not force more fluid from the tissues into the pipette, but, instead, an increased intake resulted. These find-



FIGURES 10a and 10b. Venous obstruction produces intermittent backflow from the tissues to the reservoir. Reactive hyperemia following release of the obstruction reverses the flow so that much more fluid enters the skin in spite of dilatation of the vessels. The findings are shown in FIGURE 10a as read at ½ minute intervals, and in FIGURE 10b as read at 1 minute intervals.



FIGURES 11 and 12. The skin of animals recently killed fails to take up Locke's solution. When forced into the tissue, the fluid moves through it in a continuous manner.



FIGURES 13 and 14. Relatively unabsorbable fluids, introduced at atmospheric pressure into living skin are not taken up by it. Forced into the skin under pressure, they enter continuously.

ings indicated, but did not prove, that we were dealing with an intermittent absorption of the introduced fluid, rather than with a simple intermittent movement through the tissues.

The Influence of the Circulation

If the intermittent entrance of fluid into skin in these experiments was caused by periodic absorption into the blood, then, cessation of the circulation should inhibit the phenomenon. As FIGURES 11 and 12 show, such was found to be the case. Locke's solution, brought at atmospheric pressure into contact with the dermal tissues of the ears of mice, 30 minutes to 1 hour after the animals had been killed with ether, failed, as a rule, to enter the tissues. Later, in each experiment, a small amount of pressure was put upon the introduced fluid, to force it into the tissues at rates of flow like those occurring spontaneously into normal skin at atmospheric pressure. In the absence of the circulation, as the figures show, the fluid movement through the apparatus was continuous and irregular, as seen in the control experiments. There was no intermittency like that which occurred when the same solution was forced into the same tissue in living animals with intact circulation (FIGURES 3 and 4).

The Movement of Relatively Unabsorbable Fluids through Living Tissues

Next, to determine whether the intermittency of fluid take-up was brought about by periodic absorption of fluid or merely by cyclic movement through tissues, two relatively unabsorbable fluids were brought into contact with the tissues. These fluids, homologous serum which is slowly absorbed from the tissues, and a mixture of Locke's solution with ½ per cent of a colloidal vital dye, pontamine sky blue, which produces edema and increases in bulk within the tissues, were not taken up by the skin at atmospheric pressure. Forced into the tissues under light pressure, the movement was not intermittent, but irregularly continuous. FIGURE 13 shows the findings in an experiment made with homologous serum, while FIGURE 14 shows those from a test made with the mixture of Locke's solution and dye. As the chart shows, edema developed in the latter, and, at the 16th minute, fluid ran back into the pipette. When pressure was applied, it moved into the tissues continuously.

Some Findings of Others in Relation to the Observations Just Described

Can there be, superimposed upon the continuous exchange between the blood and tissues, an intermittent preponderance, now of escape of fluid, now of resorption? Is it possible that the passage of fluid between the blood and tissues may, now in this spot, now in that, be entirely in one direction, for short periods of time? There are facts which support this supposition.

Blood flow in some tissues is known to be intermittent. Richards and Schmidt[20] noted an intermittent flow through kidney glomeruli. Krogh[21] has reported that the capillaries supplying local regions of the tongue and skeletal muscles of the frog may have blood coursing through them at one moment and be completely closed at another. Knisely[22, 23] has shown that the supply of blood to certain regions of the spleen is periodic. Grant,[24] observing irregularities in blood flow in the rabbit's ear, has discussed these in relation to the possible functioning of arteriovenous anastomoses, and has shown that the latter open and close intermittently. The recent studies of Zweifach[25-29] and of Chambers and Zweifach[30-32] on the circulation in many organs of various animals, taken with the data presented by them at this conference, have emphasized the intermittency of blood flow to localized regions through the phenomenon they term vasomotion. Zweifach's perfusion studies with dyes and particulate matter[26-30] have shown, further, that, at times, fluids may escape all along the course of a capillary and, at other times, may enter all along it.

It is well known, too, that capillary pressure varies, from time to time, in any one region, depending upon local anatomical and functional differences in the circulation.[18] Landis[18] has noted that, in an entire capillary or even a whole network of capillaries, the hydrostatic pressure may vary enormously above or below the colloid osmotic pressure of the blood. Such a change, occurring in a capillary network, following, perhaps, the intermittent opening or closing of an arteriovenous anastomosis, or as the result of the vasomotion described by Chambers and Zweifach, could account for the periodic entrance of isotonic fluid into the skin of our experimental animals, and might well determine whether fluid passed outward or inward, under normal circumstances.

More recently, Neumann, Cohn, and Burch[33-35] have reported intermittent volume changes, measured plethysmographically, occurring in the tips of fingers, toes, and ears of human subjects. They believe that

the spontaneous fluctuations in volume possibly aid in the distribution and mixing of extravascular fluids and may facilitate the passage of fluid through blood capillary and lymphatic walls. The time relationships of some of these intermittent changes, which seem to involve accumulation or loss of either extracellular fluid or fluid transported by blood or lymphatics, are much like those found in the experiments reported above on the intermittent inflow of Locke's solution into the skin.

INTRADERMAL INTERSTITIAL RESISTANCE AND EXTRAVASCULAR PRESSURE CONDITIONS WITHIN CUTANEOUS TISSUES

Knowledge of the pressures existing within tissues is essential to an understanding of lymph formation and fluid exchange. Pressure within a tissue, when it exists, must tend to decrease the escape of fluid from the blood, and affect lymph flow.

In the past, when workers attempted to determine the pressure within cutaneous tissues, the minimum pressure required to force small amounts of fluid into the normal skin was taken as implying the existence of a nearly equivalent tissue tension or tissue pressure. Few authors have realized that the amounts of fluid employed have been great enough, as a rule, to force the tissue elements apart and to set up artificial pressures. When freely movable fluid has collected in edematous skin, the existing pressure of the edema fluid can be measured directly, and the pressure is, of course, equal to the prevailing tissue pressure. Since this pressure is exerted by interstitial edema fluid upon the outer walls of blood and lymphatic capillaries, it should be distinguished from tensions or pressures within formed elements, themselves. Therefore, for the sake of clarity, it will be termed here and in later papers, *interstitial pressure*. In normal connective tissue in which there is not enough free fluid to make direct manometric readings of its pressure, the interstitial pressure cannot be directly measured. None the less, it must be exerted on the extravascular fluid, in whatever form it exists interstitially. The nearest one can come to measuring the interstitial pressure in normal skin, is to introduce into the tissue the least possible amount of an unabsorbable fluid that will serve as an indicator, and then determine the lowest pressure that will cause the slightest measurable movement inwards against the resistance of the tissues. To avoid the creation of artificial pressure, the movement should be so slow that distortion of structure is minimal or absent.

When this is the case, the pressure required to overcome the interstitial resistance should not differ much from the true interstitial pressure. Without measuring the latter directly, one should be able to estimate it with sufficient accuracy for practical purposes.

Since the method described in the first section of this paper permitted the introduction of very minute amounts of fluid into the tissues, it seemed to offer an opportunity to approach these criteria and to study the pressure conditions within tissues with somewhat greater accuracy than had been previously possible. Accordingly, studies of this sort were undertaken in the following way.

The Determination of Intradermal Interstitial Resistance

To study the pressure conditions within the tissues, the two relatively unabsorbable test fluids mentioned above were employed. It is to be recalled that one of these, a mixture of Locke's solution and dye, is a mildly edema-forming mixture, which increases slightly in bulk within the tissues[1-3]; the other, homologous serum, is slowly absorbed from tissues. The viscosities of these fluids are not unlike that of Locke's solution. One or the other of the test fluids was brought, with no pressure, as already described,[1-3] into contact with the dermal tissue of the ears, backs, or thighs of mice and rabbits anesthetized with luminal or nembutal. The meniscus of the test fluid in the injecting pipette was watched, and in approximately all instances no movement of it occurred. After a period of 10 to 15 minutes, a pressure of 0.5 cm. of water was brought to bear upon the fluid in the pipette, and, thereafter, the pressure was raised every few minutes by small increments, until flow began. Just enough pressure was then maintained to keep up an inflow of at least 0.04 c.mm. per 5 minutes and not more than 0.08 c.mm., which, it is to be stressed, corresponds to the rate at which Locke's solution enters the same tissue at atmospheric pressure; that is to say, without forcible distention of the formed elements. This minute amount of fluid introduced into the tissues sufficed for direct manometric pressure determination.

The pressure required to obtain this flow will be termed the *interstitial resistance.* It measures, of course, not only the interstitial pressure, but also the pressure necessary to overcome the resistance of the skin to the continuous passage of a relatively unabsorbable fluid, flowing interstitially at the same average rate at which an absorbable fluid, introduced into the same tissue in the same way, moves interstitially but without pressure. The data, to be presented in full in a

later paper,[36] will show that the measurements of interstitial resistance, though not actual measurements of interstitial pressure, are very close to the latter and always a little higher. For practical purposes, the difference, under the conditions of the experiments to be described,

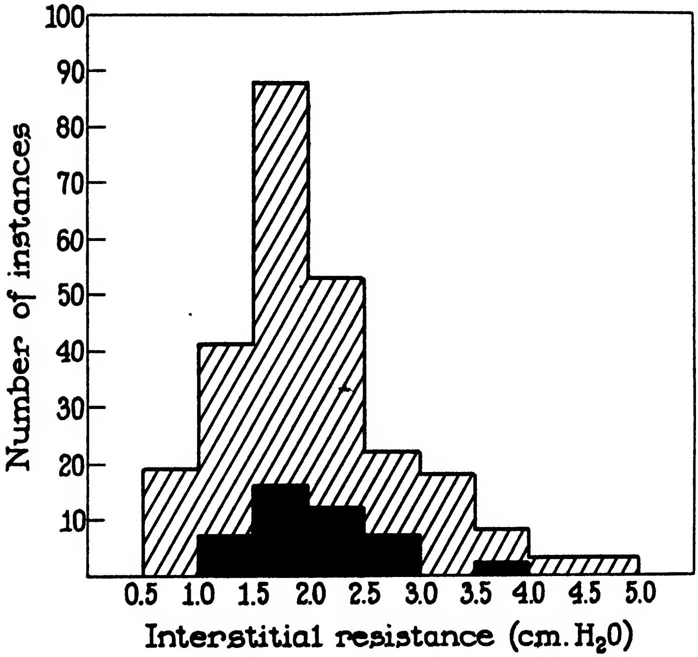


FIGURE 15. The intradermal interstitial resistance in normal skin of the ears, backs, and thighs of mice.

appears to be negligible. A discussion of the matter, together with a review of previous work, will appear in papers now in press.[36, 37]

The Interstitial Resistance in the Skin of the Mouse. The results of more than 250 determinations of the interstitial resistance made in the normal skin of the ears, backs, or thighs of mice anesthetized with luminal or nembutal are graphically presented in FIGURE 15. The cross-hatched columns indicate the number of instances in which the intradermal interstitial resistance was found between 0.5 and 1.0, 1.0 and 1.5, 1.5 and 2.0 cm. of water, respectively, and so on up to 5.0 cm. of water, the highest readings obtained in normal skin. In these instances, the measurements were made with the dye-Locke's solution.

By far the greatest number fell between 1.5 and 2.0 cm. of water, which can be taken as the normal interstitial resistance in loose skin of the mouse. Only rarely were resistances found as high as 4.5 to 5.0 cm. of water. The black columns in FIGURE 15 represent the results of 44 determinations of interstitial resistance made with homologous serum, instead of with the dye-Locke's solution.

No obvious differences were found in the interstitial resistance of the various regions of the skin. This was to be expected, for in the areas chosen for experiment the skin was loose. In work to be reported later, we have found a higher interstitial resistance in the tense, tough skin near the ankles and wrists of these animals.

CHANGES IN THE INTRADERMAL INTERSTITIAL RESISTANCE OCCURRING IN VARIOUS PHYSIOLOGICAL AND PATHOLOGICAL CONDITIONS

Measurements were next carried out in skin subjected to various physiological and pathological changes.

Hyperemia. The intradermal interstitial resistance remained practically unchanged in states of hyperemia.

Edema. In many experiments, the intracutaneous interstitial resistance was determined in normal skin which was then rendered edematous by the application of irritant chemicals. In other experiments, potent edema-forming fluids were introduced into the skin directly through the injecting apparatus. When edema formed, it became a simple matter to determine the pressure of the edema fluid, which is equal to true interstitial pressure, by applying just enough pressure to the pipette of the injecting apparatus to prevent backflow into it. In all instances in which edema appeared, the edema fluid pressure was measured first, and then the relatively unabsorbable test fluid was forced into the tissues at the proper rate, to determine the interstitial resistance of the edematous skin. Later, in the same animals, as edema formed or changed, similar determinations were made at various intervals. The circumstances usually permitted the measurement of both the edema fluid pressure and the interstitial resistance, during the formation of edema, during its height, and during the process of absorption.

The Relationship of Interstitial Resistance to Interstitial Pressure (Edema Fluid Pressure). In these experiments and in scores of others in which edema was present in skin, the interstitial resistance was found about 0.5 cm. of water higher than the edema fluid pressure. It does

not follow that in normal skin the relationship between true interstitial pressure and interstitial resistance would be the same. One can only say that the true interstitial pressure must be slightly less than the interstitial resistance.

Findings in Edematous Skin. The intradermal interstitial resistance increased but little when edema formed slowly, rising from the average of 1.7 cm. of water to 4.5 or 8.5. By contrast, when edema formed rapidly, both the edema fluid pressure and the intradermal interstitial resistance rose in proportion to the rapidity with which edema formed, reaching maxima of 10 to 15 cm. of water. As the edema subsided, the pressures fell. Later, if inflammation and induration followed, the interstitial resistance became high again, but, as the reaction subsided and the tissues became boggy, the interstitial resistance fell and often reached the normal levels, even in the presence of edema.

In Skin Inflamed or Indurated Following Trauma. In mouse skin injured by trauma, some edema usually occurred, and the intradermal interstitial pressure (edema fluid pressure) often rose to levels of 10 to 15 cm. of water within a few hours. In all these instances, the intradermal interstitial resistance was found about 0.5 cm. of water higher than the interstitial pressure. If the injury progressed to induration, the interstitial resistance rose to such high levels, 29.0 or even 33.0 and 37.0 cm. of water, that fluid could hardly continue to escape from the capillaries. In such instances, necrosis of the skin developed later. In many of the instances showing a severe reaction, no free edema fluid flowed into the apparatus at atmospheric pressure, even from ears apparently edematous. In less severe reactions, free edema fluid was usually demonstrable.

The Effects of Venous Obstruction in Mice. Following venous obstruction from a cuff placed about the thighs of mice, the interstitial resistance in the tense skin of the dorsum of the ankle showed immediate and profound changes, rising in about half an hour from an average of 2.8–4.6 cm. of water to 25.0–33.0 cm. of water. Since the skin in this region is quite tense, the pressure changes may have been greater than those which would have been obtained in other skin areas. However this may be, adequate controls showed that the changes were not due to tension in the skin brought about by the cuff. It is to be stressed that these findings, obtained by direct measurements, correspond very closely to the figure (36.0 cm. of water) established by the indirect plethysmographic methods of Landis *et al.*,[38] and Landis and Gibbon,[39] as the effective counter-pressure exerted against filtration of fluid from the blood by tissue pressure, in conditions of venous obstruction.

INTERSTITIAL RESISTANCE IN LIVING HUMAN SKIN

Since the methods employed for the preceding work utilized much less fluid than any of the techniques described by previous workers for studying intradermal pressure conditions, a few determinations were made on living human skin. To accomplish this, the apparatus was modified, as will be described in a later paper,[37] by discarding the constant temperature bath, placing the injecting pipette in a series of celluloid boxes, instead, and keeping the room temperature constant. Extensive studies were not undertaken, since the determinations by this method require such long periods of complete immobility on the part of the subject that they can be carried out only with difficulty.

On the volar surfaces of human forearms, intradermal interstitial resistances of 2.5, 2.9, 3.0, 3.1, 3.2, 4.5, and 5.1 cm. of water were encountered, and, on the dorsal surfaces, resistances of 2.1 and 2.5 cm. of water. On the dorsum of the ankle, with the subject lying flat, resistances of 2.5, 2.9, 3.0, 3.1, and 3.5 cm. of water were obtained.

Variations of Interstitial Resistance in Human Skin

The Effects of Standing. Wells, Youmans, and Miller[40] have reported an intradermal tissue pressure of 12.5 cm. of water in the legs of subjects who had been standing for 100 minutes. Burch and Sodeman[41, 42] reported a subcutaneous pressure of 10.2 cm. of water after 1 hour of standing. In the present work, we found the interstitial resistance increasing from 2.9 and 3.5 cm. of water to 24.0 and 28.0 respectively, in the skin of the ankles of men, after approximately half an hour of standing. The increase would seem to be great enough to interfere with fluid filtration from the blood. The relationship of these findings to those reported by Landis and Gibbon *et al.*,[38, 39] obtained by indirect plethysmographic methods, is obvious.

The Effects of Venous Obstruction. In a few experiments, as soon as the intradermal interstitial resistance had been measured in the skin of the ankles, or the volar surface of the forearms of human subjects, venous obstruction in the parts was produced by inflation of a pressure cuff about the thigh or about the upper arm, respectively. Shortly thereafter, backflow of fluid occurred from the tissues into the injecting apparatus. Clearly, free interstitial fluid had made its appearance. Next, the pressure in the injecting apparatus was raised, to measure the extravascular fluid pressure (true interstitial pressure), and, having found it, we forced the test fluid into the tissues at the regular rate, to obtain the interstitial resistance. Again, in these tests, the

latter was found about 0.5 cm. of water higher than the former, which increased to levels of 15.0 to 23.0 cm. of water, during 15 to 25 minutes of obstruction. The observed increases were not as great as those found in the tense skin of the ankle of the mouse, nor as great as one would expect from the earlier work of Landis and Gibbon.[39] Possible reasons for the difference will be discussed in a later paper.[44] Since, in man, the interstitial pressure rose so rapidly in the first few minutes of venous obstruction, findings similar to those in mice might have been obtained, if the tests could have been continued for longer periods. Unfortunately, complete immobility of the subjects' limbs could not be maintained for more than 20 minutes to a little over half an hour, perhaps because of the added discomfort of the pressure cuff. Nevertheless, even in these short periods of time, the interstitial pressure and resistance rose to levels high enough to reduce greatly the further escape of fluid from the blood. In the absence of factors preventing lymph movement, as for example, occlusion of the channels or the maintenance of a dependent posture in a limb, such pressures would enhance lymph formation and flow.

Discussion of the Findings

The findings here reported, taken with those of others to be reviewed in a later paper,[37] indicate that in normal, unstretched skin the interstitial resistance is low. It follows that the interstitial pressure is slightly lower, on the average, less than 1.7 cm. of water in the skin of the mouse and less than 3.1 cm. in human skin. One may conclude that, under normal conditions, interstitial pressure must exert only slight hindrance to the escape of fluid from the blood and only slight assistance to the return of interstitial fluid to it or to the lymph. By contrast, under conditions of venous obstruction and those of rapidly forming tense edema, following injury, the edema fluid pressure (in the skin of the mouse) may rise to such an extent as to constitute an effective opposition to the filtration of fluid from the vessels and to be of great aid to the formation of lymph.

It was of much interest to find, quite frequently, in the studies on inflammatory edema in mouse skin, boggy edematous tissue in which neither the edema fluid pressure nor the interstitial resistance was increased. In these tissues, there must have been some readjustment of the fixed elements, to permit the accumulation of fluid without an increase in the edema fluid pressure.

Of still greater interest was the occasional finding of boggy edematous

skin, in appearance like ordinary edematous skin, but in which no free fluid was demonstrable. In these instances, there must have been swelling or imbibition of tissue fluid by the tissue elements. This state of affairs deserves further study.

In the legs of normal persons, quietly standing, and in the arms or legs, following venous obstruction, the intradermal interstitial pressure rises to levels which must greatly hinder the escape of fluid from the capillaries.

INTRALYMPHATIC CAPILLARY PRESSURES

What effects do these changes in the interstitial pressure, exerted extravascularly upon lymphatics, have on lymph formation and flow; and what are the relationships of the pressures existing within, and just outside of, the lymphatic vessels?

Using a slight modification of the classical micromethods of Landis[43] for measuring the pressure within blood capillaries, and the techniques already described for the determination of the interstitial resistance and edema fluid pressures, some studies of the pressures existing inside and outside of the lymphatic capillaries have now been made in the ears of mice and small rabbits. Since the work, interrupted several years ago, is incomplete, the methods and findings will be fully reported later.[44] In brief, the pressures of lymph in the lymphatic capillaries are low, ranging, in normal skin, from 0.0 to 2.7 cm. of water. By far the greatest number of determinations fell between 0.8 and 1.5 cm. of water. Invariably, after the original determination of intralymphatic capillary pressure had been made, an additional pressure of 0.5 cm. of water thrown into the injecting apparatus produced flow in the channels.

Pressures Existing Outside the Lymphatic Capillary Wall. In most of these experiments, either immediately before or just after measuring the intralymphatic capillary pressure, the interstitial resistance was determined in the cutaneous connective tissue close to the vessel. In three-quarters of the instances, the interstitial resistance was found higher, by 0.5 to 1.5 cm. of water, than the pressure within the lymphatic capillaries. In the remainder of the instances, the differences were less. The interstitial resistance was never found lower than the intralymphatic capillary pressure. In these experiments on normal skin, the true interstitial pressure outside the lymphatic capillary wall could not be exactly determined, since only the interstitial resistance could be measured. However, the work already described has indi-

cated that, in skin edematous from chemical or other injury, also in skin containing free fluid during venous obstruction, the interstitial resistance is only 0.5 cm. of water higher than the true interstitial pressure, which, under these conditions, can be measured as the edema fluid pressure. If a similar relationship holds between true interstitial pressure and interstitial resistance in normal skin, then, since the interstitial resistance is usually from 0.5 to 1.5 cm. of water higher than the intralymphatic capillary pressure and never lower than the latter, there would seem to have been, in the majority of the experiments, a slight gradient of pressure between the interstitial fluid and the fluid within the lymphatic capillaries. In the remainder of the instances, the interstitial pressure and the intralymphatic pressure would seem to have been about equal.

Whatever may be the state of affairs in normal skin, in that of mouse ears rendered edematous by chemical irritants or by mild crushing injuries, the edema fluid pressure (true interstitial pressure) usually stood 1.0 to 5.9 cm. of water higher than the intralymphatic capillary pressure. Under these circumstances, a definite gradient of pressure existed between the tissue fluid and the lymph, tending to increase the latter. These measurements afford an obvious explanation for the well-known increase in lymph formation and flow in certain edematous states. Only one further point requires emphasis. It is well known, from previous studies from this laboratory[4, 8] and others,[45, 46] that, even in tense edema of cardiac disease[14] or of inflammation[12] in which the edema fluid is under considerable pressure, the lymphatic capillaries are not squeezed shut by the pressure, but, instead, are open and full of lymph, the fibrillary structure of their walls tending to hold them open. Under these conditions, a gradient of pressure from the extravascular fluid to the lymph can become effective in producing lymph flow.

LYMPHATIC PARTICIPATION IN CUTANEOUS PHENOMENA

The foregoing consideration of certain mechanical factors influencing the formation of cutaneous lymph affords a background for a review of some older observations on the behavior of lymph after its entry into the lymphatics, and of the responses of the channels, themselves, in various physiological and pathological conditions. Discussions of the physiology of the skin usually ignore the presence of the lymphatics and consider only the happenings in blood vessels, nerves, and the

interstitial fluid. This state of affairs is largely due to our ignorance, which springs from the lack of suitable means for the study of cutaneous lymphatics.

A Method for Rendering Lymphatics Visible in Living Skin of Men and Animals

Several years ago, in collaboration with Dr. Stephen Hudack, a technique was devised[47] by which the lymphatic capillaries of living human and animal skin were rendered visible. With a dissecting needle, ground as finely as possible, a minute tunnel a few millimeters in length was made in the subpapillary layer of the skin and parallel to the surface. A gauge 30 or 29 needle attached to a syringe was inserted into the tunnel and deliberately moved back and forth to tear lymphatic capillaries. Small amounts of various blue vital dyes, instilled into the tunnel with the least possible pressure, entered into the torn lymphatic capillaries and rendered them visible. The deliberate attempt to open lymphatic capillaries for the acceptance of the colored fluid differentiates this technique sharply from those already outlined in which every effort was made to preserve the integrity of the vessels in order to study the movement of fluids through the tissues themselves.

The three photographs (PLATE 1, FIGURES 1a, 1b, and 1c), magnified about 4 times, illustrate the results of the introduction of dye by this technique into the skin of the volar surface of the forearm of a normal subject. The lymphatic capillaries of the superficial plexus, lying in the sub-papillary layer of the skin, have taken up the dye. As the pictures show, this plexus is far richer than has generally been supposed. The abundance of the anastomoses deserves emphasis because of the fact, to be stressed below, that every scratch or puncture penetrating the corium tears open superficial lymphatic capillaries, allowing foreign material to enter them directly. Indeed, the lymphatic plexus is so close-meshed that one cannot make an intradermal injection without injecting lymphatics. PLATE 1, FIGURES 1a and 1b, taken about 30 seconds and 5 minutes, respectively, after the beginning of the introduction of dye, which lasted about a minute, shows these features well. In PLATE 1, FIGURE 1b, at the periphery of the injected area, one can see in one or two places pale tufts at the ends of intensely black capillaries, where the dye solution in the lymphatics was being diluted with lymph.

The Phenomenon of Streamer Formation Following the Introduction of Dye into Living Skin

Invariably, following the introduction of dye into the living skin of the arm or the ankle, pigment, pale because diluted with lymph, drained from these tufts into deeper subcutaneous channels and ap-

peared under the skin like dimly visible colored streamers ascending the limb. In PLATE 1, FIGURE 1c, the periphery of the same area photographed in PLATE 1, FIGURES 1a and 1b, is shown as it appeared after half an hour. A pointer indicates the streamer which had formed. PLATE 1, FIGURE 2, a reduced photograph of the forearm of a normal subject, shows streamers which arose from two injections and extended up the arm.

The occurrence of such streamers, following the intradermal introduction of dye, indicated that they might be used to study changes occurring in lymph flow. After some experimentation, it was found that very small volumes, 0.01 to 0.03 cc. of dilute, 1 per cent, dye solutions yielded colored streamers which formed so slowly that variations in their length and intensity could easily be distinguished.

It is to be stressed that the development of a streamer, following the introduction of a minute amount of dye, is a very different phenomenon from that which occurs when larger amounts of dye are forcibly injected into skin, as in making an ordinary clinical intradermal injection. Then, undiluted dye is actually forced into the lymphatics under high pressure and is not transported by flow along them. By contrast, in the technique just described, the presence of the pre-formed tunnel permits one to instil into it, with the least possible pressure, minute amounts of dye solutions, isotonic with blood. From the injected area, the solution, diluted by tissue fluids, extends very slowly into superficial lymphatics, there to be still more diluted. Slowly, and only after some minutes, it reaches the lymphatic trunks, to become still further diluted by lymph coming from other areas of the skin. To be sure, some pressure is unavoidably employed in introducing the dye, but it is slight indeed. Recently, we have had opportunities to measure it. At the needle's tip, the pressure varies from 6 to 12 cm. of water and, in the tunnel, from 4 to 8 cm. of water. The pressure falls rapidly, to become equal, within 4 minutes, to the usual interstitial resistance of less than 2.0 cm. of water. Nevertheless, one may well ask: Does the extension of the streamer developing in a normal limb approximate the movement of lymph taking place naturally, or is there an artificial movement caused by local edema at the site of injection?

The Relationship of Streamer Formation to Lymph Movement. To test the point, lymph was collected in many experiments from lymphatics at the base of one ear of large rabbits. In those instances in which lymph flow became sufficiently constant after it had been collected for some time, intradermal injections of 0.01 to 0.03 cc. of dye

solution were made, in the usual manner, at the tips of both ears. Color appeared in the lymphatics near the injections and passed along the channels to each ear base, where, on the cannulated side, it entered the cannula. In some of the experiments, during this period, the rate of flow of collected lymph did not increase; in some, the flow increased by 10–15 per cent. Rarely, an increase of 20 per cent, and rarely, too, a decrease was noted. The time intervals required for the blue streamers of dye to reach the bases of both ears were approximately the same, 25–35 minutes.

Since the streamers in the cannulated ears reached the cannulae at their bases at the same time that they reached the bases of the normal ears, and since the volume of lymph collected in the cannulae before and after injection did not increase greatly, and since the streamers were composed of dye much diluted by clear lymph, it follows that the rate of movement of the streamers gave a fair approximation of the rate of lymph movement from uninjected tissues. This conclusion was confirmed in another way.

The Effect of an Increase in the Rate of Lymph Flow upon the Rate of Streamer Formation. It is well known that an application of heat or the formation of edema increases the rate of lymph flow.[46, 47, 48] Accordingly, experiments were made to test the effects of an increase in lymph flow upon the rate at which the blue streamers moved. Again, lymph was collected from the base of one ear of several large rabbits. In the instances in which flow was sufficiently constant for some time, both ears were warmed by heated air or painted with xylol to induce edema and hyperemia. Dye was then injected at the tips of both ears, in the usual manner and amount. The resulting colored streamers moved up the ears to their bases in less than half the time required in the preceding experiments, and again blue fluid entered the cannulae. The volume of lymph collected per 5 minutes in the cannulae, following the application of heat or xylol, was doubled or even quadrupled, and was greatest in those instances in which the movement of the streamers was most rapid.

Changes in the Length and Intensity of Colored Streamers of Human Skin Accompany Conditions Known to Stimulate or Retard Lymph Flow. It is common knowledge that lymph flow is increased by massage,[46, 49] by activity,[50, 51] by hyperemia,[52] and, as already mentioned, by applications of heat[46, 47, 48]; further, that it is diminished in limbs at rest.[44, 48, 53] Constant amounts of a standard dye solution were injected into corresponding skin areas of both arms or legs of normal

human subjects. One limb was then kept at rest, which is known to diminish lymph flow; the second limb, in some subjects, was kept at rest but heated, or in other tests massaged or used for muscular exercise. In every instance, the limbs at rest showed short, pale streamers, which developed very slowly, while the limbs heated, massaged, or used to pummel a punching bag showed rapidly developing, long, and deeply colored streamers. In limbs injected and then elevated, streamers developed rapidly; in those injected and allowed to hang downwards, none or almost none appeared.

CHANGES IN LYMPH FLOW IN CERTAIN PATHOLOGICAL CONDITIONS

With the fact established, that variations in the length and rapidity of streamer formation in human skin accompanied conditions known to stimulate or retard lymph flow, the method was used to indicate changes in certain conditions having an unknown effect upon it. The present paper will review only a few findings of clinical interest, taken from a number of others, already described elsewhere.[13, 14]

Streamer Formation Following the Release of Venous Obstruction. It is of much interest that exceedingly rapid streamer formation appeared in resting, horizontally placed limbs, during the intense reactive hyperemia which followed the release of venous obstruction. For example, in a test, during which the photographs shown in PLATE 1, FIGURES 3a and 3b, were taken, a pressure cuff on the upper arm was inflated to a pressure of 90 mm. of mercury, for 25 minutes. As illustrated in FIGURE 3a (magnified about 1½ times), dye introduced into the lymphatic capillaries of the volar surface of the forearm, during the first 5 minutes of the obstruction, showed no tendency toward streamer formation. Almost immediately following the release of obstruction, an intense reactive hyperemia developed, and a photograph, PLATE 1, FIGURE 3b, taken only 2 minutes after its appearance, shows part of the streamer that had formed in this short time. In this test and in others like it, streamers reached an intensity and length never equaled in normal resting arms, and appeared like those which developed in arms injected with dye and then used in violent muscular activity.

Clearly, very rapid lymph movement occurs in resting arms, during the intense reactive hyperemia that follows the release of venous obstruction. Findings obtained with the injecting apparatus, and described in the earlier sections of this paper, offer an explanation for

such a result. It is to be recalled that, both in the ear of the rabbit and in human skin, venous obstruction led to the output of free fluid into the tissues, as shown by the fact that fluid ran back into the injecting apparatus against atmospheric pressure. Further, in human skin, after 15 to 25 minutes of venous obstruction by a pressure cuff, the interstitial fluid pressure, that is to say, the pressure of the fluid which had collected in the tissues, had risen to values between 15 and 23 cm. of water. Such pressures would tend to increase lymph flow greatly, following the release of the pressure cuff. Moreover, data like those presented in FIGURES 10a and 10b indicated that, during the reactive hyperemia following the release of venous obstruction, an increased take-up of extravascular fluid occurred. There is no reason to believe that the lymphatics did not share in the take-up.

The Presence or Absence of Lymph Movement in Edematous Skin

Interesting findings followed a comparison of the behavior of dye introduced into the edematous skin of the ankles of aged patients suffering from cardiac decompensation and poor peripheral blood flow, with that occurring in youthful nephrotic patients with good circulation Some of the findings are illustrated in PLATE 2, FIGURES 4a, 5a, and 6a, which, like PLATE 1, FIGURE 3a, have been magnified about 1½ times, instead of 4 times, as in PLATE 1, FIGURES 1a, 1b, and 1c. Since, in these tests, dye was introduced into the skin of the ankle, PLATE 2, FIGURE 4a has been presented for comparison with the others, to show the appearance of dye in the lymphatic capillaries of a normal subject's ankle The reduced photograph, in PLATE 2, FIGURE 4b, taken after 20 minutes, shows the short double streamer that had developed from this injection, while the leg remained at rest in a horizontal position. By contrast, PLATE 2, FIGURE 5a demonstrates the appearance of the lymphatic capillaries, at the same magnification, after the introduction of the same amount of dye into the swollen, edematous ankle of a cardiac patient who was losing his edema rapidly. The lymphatic capillaries were dilated and full of fluid. Dye entered them so rapidly that a much wider area of the lymphatic plexus became injected. In PLATE 2, FIGURE 5b, the leg is seen as it appeared after half an hour. In all of these patients, in spite of the larger amount of dye reaching the lymphatic capillaries, no streamers ever developed in legs that remained horizontal and at rest, even when, as in this instance, the patients were rapidly losing their edema.

Although there was no evidence of lymph movement in these patients,

it could readily be shown that the lymphatics were open. Injections in the edematous ankles, followed by elevation of the limb or by massage, promptly led to the development of streamers.

In considering the factors responsible for this stagnation of lymph, it seemed possible that the high venous pressure usually present during cardiac decompensation might retard or obstruct the escape of lymph from the larger lymphatics into the veins at the neck. Consequently, a few patients were placed in bed with their legs extended horizontally and in such a position that the skin of the forearm and that of the ankles lay at the same level, below the clavicle. After an hour in this position, dye was introduced into the skin of the arms and ankles. Streamers developed in the arms which were not edematous, but failed to appear in the edematous legs. The findings seemed to rule out the influence of high venous pressure. In some of these patients, the lymphatics seemed to be so dilated that the valves were probably no longer efficient. In such instances, dye introduced into the edematous skin of the ankles and massaged toward the toes could be seen running in that direction. Retrograde movement of this sort was never seen in healthy persons. In passing, it may be added that no streamers were seen in studies made on five patients suffering from idiopathic lymphedema.

In contrast to these findings, the youthful patients suffering from nephroses showed, during the early stages of the development of edema, a slightly greater tendency toward streamer formation than that observed in normal skin. In all, the onset of diuresis was accompanied by the development of extraordinary streamers. PLATE 2, FIGURES 6a and 6b, illustrates the findings in this condition. The patient, a young man, was injected repeatedly, over a period of weeks, as his edema slowly increased. At first, the appearance of the lymphatics and the streamers differed so little from the usual that no comments need to be made. Finally, at the height of his edema, the lymphatic capillaries became much dilated, as shown in PLATE 2, FIGURE 6a, appearing almost like those seen in the patients with cardiac disease, but with one marked difference: dye in the lymphatic capillaries at the edge of the injected region became diluted rapidly and began to drain away visibly in a very short time. In the figure, this can be seen in the upper right corner of the photograph, which was taken only 45 seconds after the beginning of the injection.

By chance, this test occurred on the day that the patient developed spontaneous diuresis. Within 8 minutes, long streamers had developed

in the resting horizontal leg,[14] and, in 25 minutes, they appeared as seen in PLATE 2, FIGURE 6b. Shortly thereafter, the streamers reached the level of Poupart's ligament. Such speed and intensity of streamer formation was never seen in normal subjects, but was found in other patients, under similar conditions.

These findings, fully described elsewhere,[4, 13, 14, 47] have been briefly sketched here to indicate that, under conditions in which free fluid is present in the skin and the circulation is good, as during the reactive hyperemia following the release of venous obstruction and in the edema of nephroses in patients with good cardiac action, lymph movement may be very great, even in motionless limbs. In the edema accompanying cardiac disease, with failing circulation, there seems to be no movement of lymph in resting limbs.

In this connection, it is of interest that,[4, 5, 47] in rabbits' ears perfused, on the one hand, with a pulsating stream of blood and, on the other, with a non-pulsatile flow, streamers rapidly became intense and long, when the perfusion of blood was pulsatile, but showed little or no tendency to form, when the flow was non-pulsatile. Apparently, the pulsation of blood vessels increases the formation and flow of lymph.

SKIN LYMPHATICS AND THE DEFENSE OF THE BODY AGAINST INJURY AND INFLAMMATION

Since the lymphatic capillaries in the superficial layers of the skin were found to be so abundant, studies were undertaken to determine what part they might play in the reactions of cutaneous tissues to injuries of various kinds. Changes in lymphatics accompanying inflammation will not be discussed, for they have been studied by Menkin,[57] who has made them the subject of the following paper.

The Behavior of Skin Lymphatics In and About Mild Burns In collaboration with Hudack, tests were carried out in mildly burned areas of animal[9, 10, 12] and human skin.[11] When sharply localized, standardized, mild burns were made in the mid-portion of the ears of mice, dye solutions introduced into lymphatics at the ear margins under slight pressure, not too long after the burns had been made, often passed in lymphatics directly through or just under the burns and appeared in normal channels beyond. When this happened, the lymphatics seemed to pour their colored contents directly into and about the area of injury, so fast that one might doubt the existence of lymphatic walls. However, in similar experiments, injections of India ink indicated the persistence of the lymphatics' anatomical continuity, for the particulate

matter failed to escape from the channels. When dye or ink injections were made at the ear margins, several hours after producing the burns, dye failed to pass into them, and instead flowed through lymphatics surrounding the burns. Of course, no injected fluids entered more severe burns, at any time, if the tissues had been coagulated.

During the later stages of the repair of burns, dye or India ink, introduced into the lymphatic capillaries at the ear tip, entered into a wealth of channels in and about the healing area. PLATE 2, FIGURE 7 (magnified about 2½ times), shows the result of such an introduction of India ink at the tip of an ear burned so severely, 9 days before, that a perforation had resulted. Rapid healing was in progress and the ink appeared in twig-like lymphatic capillaries which had grown into the tissues that were closing the perforation. Proximal to the injury, the ink entered an extremely rich plexus of newly formed channels situated in relatively healthy tissue. In other experiments, dye introduced into such lymphatics was diluted and rendered pale much more rapidly than in the lymphatics of normal skin, indicating that active take-up of fluid was occurring around the healing burns. The findings suggested a marked activity of the lymphatic system in the processes of repair in burns.

Lymphatic Participation in the Repair of Incisions. To learn something about the part played by the lymphatics in the healing of wounds and in the repair of connective tissue injuries, incisions about 1 cm. long were made in the skin of mouse ears, midway between the tips and the various bases. At varying intervals after making the incisions, dye was introduced into the lymphatics at the tips of the ears.

The behavior of lymphatics severed by incision differs greatly from that of the blood vessels,[12] for, unlike the latter, severed lymphatics may remain open for considerable periods of time and lead fluids away from wounds. The phenomenon readily explains the frequency of infection by the lymphatic route.

In some experiments, under a microscope, crystals of dye were pushed into minute intradermal puncture wounds at the margins of the ears, a few hours after making the incisions. Within 15 to 20 minutes, colored fluid could be seen passing along the channels and escaping from the severed lymphatic capillaries into the incisions, although, at this time, constriction and spasm of the blood vessels prevented all bleeding. Moreover, lymphatics proximal to the incisions took up the colored fluid and drained it away to the bases of the ears. This occurred, of course, without the application of pressure. In other experiments, dye

solutions were introduced into the lymphatics, in the usual manner, by instilling them into pre-formed tunnels at the margins of incised ears. PLATE 2, FIGURE 8 (magnified 4 times), illustrates the finding in such an experiment in which a dye solution was introduced at the ear margin, 5 hours after an incision had been made deep enough to sever both lymphatics and blood vessels. The incision is seen filled with dye and the channels appear to be draining it toward the base of the ear.

Whether or not these phenomena took place, seemed to depend upon the state of the fibrin clot and the degree of dryness in the wound. When dye was introduced into the lymphatics, a few hours later, in the usual way and with the "least possible pressure,"[12] it failed to reach the incisions, but, when subjected to pressures of 20 to 40 cm. of water, it entered them readily. Further, the slightest touch on the skin, as might be made in dressing a wound, served in these instances to force dye in lymphatics from the ear tips into the wound and on into draining lymphatics. There is good reason to suppose that the same holds true for man.

New Formation of Minute Lymphatics in Areas of Repair. Clear evidence of the new formation of lymphatics was obtained in ears studied 7 to 10 days after incision of the skin. At this time, dye solutions or India ink, introduced into the lymphatics at the periphery of the ear in the usual way, passed not only into channels surrounding the cuts, but flowed directly through the incised areas in reconstituted lymphatics. PLATE 2, FIGURE 9 (magnified 17 times), illustrates the result of an introduction of India ink at the tip of an ear incised 9 days before. The reconstituted channels appear to be carrying ink directly through the healing incision which runs almost horizontally in the photograph.

The observation—that lymphatics regenerate—is, of course, not new. The phenomenon has been described by Lee,[58] by Colin,[59] and by Reichert,[60] and much work has been devoted to the subject by Clark and Clark.[61, 62] The experiments on the mouse ear, reported here and previously,[12] were later amplified and extended by Pullinger and Florey,[45] with similar findings. These observations, like those given above, have been cited here to show that the lymphatic system is active in the processes of healing.

EVERY INTRADERMAL INJECTION IS IN PART A LYMPHATIC INJECTION

The tests on human skin, described earlier, have all indicated that the superficial cutaneous lymphatic capillaries are so abundant that intradermal injections introduce some foreign material directly into

them. Under these circumstances, skin reactions, which serve for diagnoses in certain immunological tests, such as the skin tests for allergy or the Schick and Dick tests, cannot be considered as purely local affairs. That the lymphatics are involved in local infective processes has long been known, but it is not generally recognized how readily noxious materials have immediate access to the lymphatics, once the primary barrier of the epidermis is down.

To bring out this point, several experiments were made. Sterile, sharp needles were dipped into dye solutions or suspensions of dye particles, and the skin lightly punctured with them. Under the microscope, the dye solutions or dry particles appeared in the lymphatics close to the puncture. Isotonic vital dye solutions or suspensions of dye particles placed upon superficial scarifications of the skin, like those employed for clinical vaccination, too superficial to elicit bleeding, were taken up by the lymphatics and carried away. If a knife, dipped in a dye solution or a suspension of fine particulate matter, was used to cut the skin superficially, and the cut was then sucked, as one might suck a similar injury in everyday life, the pressure thus exerted upon the skin forced the dye or the particles which had entered the torn lymphatics several centimeters up the channels. The foreign material could not be squeezed back into the cut. The experiments showed that, however slight the injury, colored particulate or diffusible matter, punctured, scratched or injected into the skin, found its way into the regional lymphatics. Further, as already described, severed lymphatics may remain open for a long time. As result of all this, the matter of local injection assumes greater importance, in the light of the fact that intradermal injections are, to a considerable extent, intralymphatic. Indeed, every local injection is, in reality, a general one.

The Transport of Foreign Substances by Way of the Lymph Is More Rapid Than the Volume of Flowing Lymph Indicates. These observations, when related to the studies on the formation of streamers, already described, bring out an important point. The volume of lymph flow represented by a colored streamer moving several centimeters up a limb is very small indeed. But skin lymphatics are small too, and smaller than they look, for they are often flattened and ribbon-like. As a result, an insignificant volume of lymph flow through a channel may serve to carry dye particles or other foreign materials, that have entered the lymph stream through a scratch or puncture, much farther through the body than one would suspect from a consideration of the volume of lymph flow alone.

THE FORMATION OF ANTIBODIES WITHIN LYMPH NODES

Since every scratch or puncture may introduce foreign materials directly into the lymphatics, why does not infection follow almost every surface injury? It is well known that lymph nodes act as filters for foreign substances entering the lymph stream, but they are not perfect filters. It seemed probable that the nodes might do more than act in this way, that they might take part in the formation of antibodies.

Several different sorts of experiments made with Hudack brought out the fact that lymph nodes, nearest the site of an intradermal injection of pathogenic bacteria, form antibodies[55] before these appear in any noteworthy amount in the blood. Many other experiments ruled out the possibility that the antibodies found in the regional lymph nodes had really been formed elsewhere in the body.

The work demonstrating these facts has already been published at length in earlier papers[55, 56] and need not be detailed again. Here we will merely suggest the way in which one type of experiment was done. Killed cultures of agglutinin-forming bacteria were intradermally injected into one ear of large numbers of mice. Into the other ears of the mice, Schick test toxin was injected. Daily thereafter, for some days, the serum, extracts from the cervical nodes of both sides, of nodes elsewhere in the body, of the liver, and of the spleen were tested for agglutinin content. After several days, agglutinins appeared, first, in high concentration, in the extracts of the cervical nodes draining the ears injected with the agglutinin-forming bacteria. They were present, too, in the blood, in traces, but they were absent from the extracts of the lymph nodes draining the ears injected with the Schick toxin and from the extracts of the other tissues. As the ears and the cervical lymph nodes on both sides were inflamed to the same extent, agglutinins formed elsewhere in the body and present in the blood would have had equal opportunity to be taken up by the cervical nodes of both sides, but they appeared only in the nodes of one side, that injected with the agglutinin-forming bacteria.

Each day, the agglutinin content increased in the extracts of the nodes from the side injected with the agglutinin-forming bacteria, and in the blood, too. However, it was not until a week later that they appeared in the extracts from the cervical nodes of the other side, or in the extracts of tissues taken elsewhere from the body.

Experiments of the same type, but made upon rabbits, were next undertaken with Kidd,[56] to learn whether the substances which neutral-

ize viruses are also formed in the lymph nodes. It was found that antiviral substances neutralizing vaccinia virus appeared first, in significant amounts, in lymph nodes nearest to the site of an intradermal injection of the virus.

SUMMARY

Methods have been devised to bring microscopic amounts of fluid into contact with cutaneous connective tissue, under pressure or without pressure, in such a manner that the fluid usually enters neither blood capillaries nor lymphatics directly. The take-up of fluid, brought into contact with the tissues in this way, has been measured and its characteristics studied.

Locke's or Tyrode's solutions, brought into contact with the cutaneous tissues by the method described and at atmospheric pressure, pass into the tissue intermittently. Forced into the skin by pressures of 1.0 to 2.0 cm. of water, the take-up is still intermittent in character. From this, it follows that either the absorption of interstitial fluid from localized regions is periodic, or the movements of interstitial fluid are influenced by intermittent physiological changes.

Hyperemia of the tissues, with accompanying dilatation of the blood vessels, increases the entrance of fluids at atmospheric pressure, but it is still intermittent. By contrast, venous obstruction leads to the appearance of free extravascular fluid in the tissues and to its intermittent backflow into the apparatus. The reactive hyperemia which follows release of the obstruction is attended by an increase of flow from the apparatus into the tissues, in spite of the great dilatation of vessels. The inflow is also intermittent.

If the skin is deprived of circulation, fluid does not enter it at all at atmospheric pressure, though it moves in regularly and continuously if slight pressure is put upon it. An edema-forming fluid, described in the text, also enters in a continuous manner when forced into the skin of either living or dead animals. So, too, does serum.

The findings indicate that, in small, localized regions of skin, the passage of interstitial fluid into the blood vessels, as well as its escape from them, may be intermittent, under normal circumstances. The occurrence of intermittent flow in the blood vessels of several tissues, observed by other workers, will go far to account for the phenomenon. It seems probable, then, that conditions of fluid exchange are constantly changing here and there in the skin, and that the passage of fluid to and from the blood is not evenly balanced in all areas at all times, but undergoes periodic fluctuations, with preponderance of inflow at certain times and of outflow at others.

The techniques described were used to investigate pressure conditions within the tissues. A quantity termed interstitial resistance, which can be measured, has been defined above and shown to be very slightly higher than the probable pressure within tissues, here termed interstitial pressure. In normal skin, as determined through measurements of interstitial resistance, the interstitial pressure, which is extravascular, is low, less than 1.7 cm. of water in the skin of mice and less than 3.1 cm. of water in human skin. On the other hand, following venous obstruction or during the rapid formation of edema, and especially following trauma or induration of the tissues, the extravascular fluid pressure, as measured directly, and the interstitial resistance may rise to such an extent as to constitute an effective opposition to the filtration of fluid from the vessels and serve as an important force in lymph formation. The interstitial resistance in human skin rises rapidly in the legs of standing subjects to levels capable of greatly hindering capillary filtration from blood vessels.

The intralymphatic capillary pressure in normal mouse skin is low, varying between 0.0 and 2.7 cm. of water. In the same tissues, the interstitial resistance outside the lymphatic capillaries is higher than the lymph pressure, in three quarters of the instances, by 0.5 to 1.5 cm. of water, and never less in the remainder. The findings suggest that there may often exist a gradient of pressure from the tissues to the lymphatic capillaries. In skin which has become edematous rapidly, the interstitial pressure, as measured by finding that of the edema fluid, is always higher than the intralymphatic capillary pressure, by pressures varying from 1.0 to 5.9 cm. of water. Such a pronounced gradient of pressure, from the tissues to the lymph, under conditions of acute inflammatory edema, readily explains the increased lymph flow occurring under these circumstances. In long standing edema, when the tissues are boggy, the edema fluid pressure is often no higher than the normal interstitial resistance. Sometimes, there is not even any demonstrable free edema fluid present in the tissue. Under these circumstances, no increase in lymph flow is to be expected, even though the channels may be full of fluid, as in the skin of patients with cardiac edema.

The experiments described in the latter part of this paper have been presented to afford a better conception of the importance of the lymphatic system in everyday life. The superficial plexus of lymphatics is so abundant that the skin cannot be scratched, cut, or injected without introducing foreign materials directly into the channels. Unlike the

blood vessels, severed lymphatics may remain open for several hours and infectious materials may readily enter them. The transport of foreign substances by way of the lymph is much more rapid than the volume of lymph flow would lead one to suspect. Often the channels are flat and ribbon-like, and, as a result, the flow of a very small amount of lymph may carry the foreign materials far through the channels.

The role of the cutaneous lymphatics in edema is of interest. In cardiac edema, perhaps because of extreme dilatation of the channels which renders their valves incompetent, the lymphatics fail in their function of fluid drainage and so add to the disability. On the other hand, in the edema accompanying nephroses, the cutaneous lymphatics aid in the drainage of fluid from the skin. During the periods of diuresis, there is an extraordinarily rapid lymph flow.

Finally, it is most noteworthy that every scratch and puncture, every injury that breaks the continuity of the skin, introduces foreign substances directly into the lymphatics. Every local injection is, in reality, a general one and, as a result of lymphatic drainage, the regional lymph nodes play their part as the first line of defense, for, in the nodes, antibodies are first formed against both bacteria and viruses.

From all this, it follows that what happens in the skin assumes greater importance, since it has now been shown that intradermal injections are partly injections into the lymphatic system, and that the immunity against disease, conferred by preventive injections, and even the reaction to the injection itself are not merely skin phenomena, but a generalized activity of the lymphatic system.

BIBLIOGRAPHY

1. **McMaster, P. D.**
   1941. J. Exp. Med. **73**: 67.
2. **McMaster, P. D.**
   1941. J. Exp. Med. **73**: 85.
3. **McMaster, P. D.**
   1941. J. Exp. Med. **74**: 9.
4. **Parsons, R. J., & P. D. McMaster**
   1938. J. Exp. Med. **68**: 353.
5. **McMaster, P. D., & R. J. Parsons**
   1938. J. Exp. Med. **68**: 377.
6. **Parsons, R. J., & P. D. McMaster**
   1938. J. Exp. Med. **68**: 869.
7. **McMaster, P. D., & R. J. Parsons**
   1939. J. Exp. Med. **69**: 247.
8. **McMaster, P. D., & R. J. Parsons**
   1939. J. Exp. Med. **69**: 265.

9. **Hudack, S. S., & P. D. McMaster**
1932. J. Exp. Med. **56**: 223.
10. **McMaster, P. D., & S. S. Hudack**
1932. J. Exp. Med. **56**: 239.
11. **Hudack, S. S., & P. D. McMaster**
1933. J. Exp. Med. **57**: 751.
12. **McMaster, P. D., & S. S. Hudack**
1934. J. Exp. Med. **60**: 479.
13. **McMaster, P. D.**
1937. J. Exp. Med. **65**: 347.
14. **McMaster, P. D.**
1937. J. Exp. Med. **65**: 373.
15. **McMaster, P. D., S. S. Hudack, & P. Rous**
1932. J. Exp. Med. **55**: 203.
16. **McMaster, P. D., & S. S. Hudack**
1932. J. Exp. Med. **55**: 417.
17. **Hudack, S. S., & P. D. McMaster**
1932. J. Exp. Med. **55**: 431.
18. **Landis, E. M.**
1934. Physiol. Rev. **14**: 404.
19. **Lewis, T.**
1927. The Blood Vessels of the Human Skin and Their Responses. Shaw & Sons, Ltd. London.
20. **Richards, A. N., & C. F. Schmidt**
1924–1925. Am. J. Physiol. **71**: 178
21. **Krogh, A.**
1929. The Anatomy and Physiology of Capillaries. Revised edition. Yale University Press. New Haven.
22. **Knisely, M. H.**
1936. Anat. Rec. **64**: 449.
23. **Knisely, M. H.**
1936. Anat. Rec. **65**: 23.
24. **Grant, R. T.**
1929–1931. Heart **15**: 281.
25. **Zweifach, B. W.**
1934. Anat. Rec. **59**: 83.
26. **Zweifach, B. W.**
1936–1937. Am. J. Anat. **60**: 473.
27. **Zweifach, B. W., & C. E. Kossmann**
1937. Am. J. Physiol. **120**: 23.
28. **Zweifach, B. W.**
1939. Anat. Rec. **73**: 475.
29. **Zweifach, B. W.**
1940. Am. J. Physiol. **130**: 512.
30. **Chambers, R., & B. W. Zweifach**
1940. J. Cell. & Comp. Physiol. **15**: 255.
31. **Chambers, R., & B. W. Zweifach**
1944. Am. J. Anat. **75**: 173.
32. **Zweifach, B. W., B. E. Lowenstein, & R. Chambers**
1944. Am. J. Physiol. **142**: 80.
33. **Neumann, C., A. E. Cohn, & C. Burch**
1942. Am. J. Physiol. **136**: 448.

34. **Neumann, C.**
1942–1943. Am. J. Physiol. **138**: 618.
35. **Neumann, C., W. T. Lhaman, & A. E. Cohn**
1944. J. Clin. Invest. **23**: 1.
36. **McMaster, P. D.**
1946. J. Exp. Med. (In press).
37. **McMaster, P. D.**
1946. J. Exp. Med. (In press).
38. **Landis, E. M., L. Jonas, M. Angevine, & W. Erb**
1932. J. Clin. Invest. **11**: 717.
39. **Landis, E. M., & J. H. Gibbon, Jr.**
1933. J. Clin. Invest. **12**: 105.
40. **Wells, H. S., J. B. Youmans, & B. G. Miller**
1938. J. Clin. Invest. **17**: 489.
41. **Burch, G. E., & W. A. Sodeman**
1937. J. Clin. Invest. **16**: 845.
42. **Burch, G. E., & W. A. Sodeman**
1937. Proc. Soc. Exp. Biol. & Med. **36**: 256.
43. **Landis, E. M.**
1931. Heart **15**: 209.
44. **McMaster, P. D.**
Unpublished data.
45. **Pullinger, B. D., & H. W. Florey**
1935. Brit. J. Exp. Path. **16**: 49.
46. **Drinker, C. K., & J. M. Yoffey**
1941. Lymphatics, Lymph, and Lymphoid Tissue. Harvard University Press. Cambridge.
47. **McMaster, P. D.**
1941–1942. The Harvey Lectures. **XXXVII**: 227.
48. **Starling, E. H.**
1909. The Fluids of the Body. A. Constable & Co. London.
49. **Lazarus-Barlow, W. S.**
1894. Phil. Trans. Roy. Soc. London B **185** (II): 779.
50. **Asher, L., & A. G. Barbara**
1898. Z. Biol. **36**: 154.
51. **Bainbridge, F. A.**
1900–1901. J. Physiol. **26**: 79.
52. **Field, M. E., C. K. Drinker, & J. C. White**
1932. J. Exp. Med. **56**: 363.
53. **Weech, A. A., E. Goettsch, & E. B. Reeves**
1934. J. Exp. Med. **60**: 63.
54. **Abramson, H. A., & M. Engel**
1938. J. Invest. Dermat. **1**: 65.
55. **McMaster, P. D., & S. S. Hudack**
1935. J. Exp. Med. **61**: 783.
56. **McMaster, P. D., & J. G. Kidd.**
1937. J. Exp. Med. **66**: 73.
57. **Menkin, V.**
1940. Dynamics of Inflammation. Macmillan. New York.
58. **Lee, F. C.**
1922. Bull. Johns Hopkins Hosp. **33**: 21.

59. **Colin, G.**
1873. Traité de physiologie comparée des animaux. **2**: 238. Second edition. J. B. Baillière & fils. Paris.

60. **Reichert, F. L.**
1926. Arch. Surg. **13**: 871.

61. **Clark, E. R., & E. L. Clark**
1932. Am. J. Anat. **51**: 49.

62. **Clark, E. R., & E. L. Clark**
1933. Am. J. Anat. **52**: 273.

DISCUSSION OF THE PAPER

Dr. Mudd:

Can anything be said about the amount of antigen which might escape the filtering action of lymph nodes, after its introduction into the peripheral lymphatics?

Dr. Drinker:

Some antigen escapes into the general circulation.

Dr. William Ehrich (*University of Pennsylvania and Philadelphia General Hospital, Philadelphia, Pa.*):

Following the injection of antigen into the foot-pad of the rabbit, the nearest lymph node, the popliteal node, becomes most enlarged, the next node in the line of flow is less enlarged, and so on, as the lymph moves toward the blood stream from the periphery.

Dr. George Burch (*School of Medicine, Tulane University, New Orleans, La.*):

The variations in the rate of flow of lymph in the superficial lymphatics of the skin, described by Dr. McMaster, and vasomotion, described previously by Drs. Chambers and Zweifach, indicate the complex nature and great lability of the peripheral blood and lymphatic vessels. By plethysmographic means, we[1] were able to record, simultaneously from the finger and toe tips and pinnae, considerable variations in volume of enclosed parts. The nature of the volume changes indicated that they were within the blood vessels of the parts studied. The pulsations noted in the lymphatics, by Dr. McMaster and by Drs. Nicoll and Webb, suggest that the variations in volume recorded by the plethysmograph are probably due, in part, to lymph vessels, as well as blood vessels.

In an attempt to understand, qualitatively and quantitatively, we divided them into five types of volume waves or, better, deflections; viz., *pulse* deflections, *respiratory* deflections, *alpha* deflections, *beta* deflections and *gamma* deflections. The *pulse* deflections are the result of each cardiac ejection; the *respiratory* deflections are the result of changes associated with the phases of respiration; the *alpha, beta,* and *gamma* deflections are related to sympathetic nervous system activity, general shifts in blood volume, and factors not too well known. The qualitative and quantitative nature of the *alpha* deflections were more clearly defined.

The complex nature of vasomotion and of the "spontaneous" variations in volume of the small blood vessels was suggested by the simultaneous plethysmographic measurements of the three widely separated parts. For example, the alpha, beta, and gamma deflections would show simultaneous, small or large shifts of blood from the three parts, evidenced by simultaneous decreases in volume of these parts. This indicated a shifting of blood from the surface of the body to the interior. Similarly, simultaneous swelling of these parts indicated a shifting of blood from the interior of the body to its surface. During some moments, a finger tip would fill with blood while a toe tip would reduce its blood volume,

[1] Burch, G. E., A. E. Cohn, & C. Neumann. Am. J. Physiol. 138: 423. 1942.

indicating local phenomena concerned with local adjustments in blood flow. Although we made no attempts to record variations in blood volume in the viscerae, simultaneously with those in the external parts, the studies of Markee on endometrical transplants in the anterior chamber of the monkey's eye, and the studies of others on the liver, spleen, and kidney, showed volume changes in the blood vessels of these organs, which were very similar to those recorded for the finger and toe tips and pinnae. These studies support the idea of blood being shifted from one portion of the body to another, to meet local and immediate demands. Such a mechanism of "lending and borrowing" of blood from one portion of the body to another makes it possible to cope with great demands for blood in any one organ, at a given moment, at the expense of another organ. Any organ may become engorged with blood, when necessary. Thus, these great local demands are fulfilled with a relatively small blood volume, about 5 liters. Without such a mechanism to meet emergency requirements for large amounts of blood, the blood volume of man would have to be several times greater. He would be a larger and heavier animal, with a larger and stronger heart and vascular system. These phenomena indicate the purposeful or less haphazard nature of the variations in volume of blood vessels, lymphatics and vasomotion. They represent physiologic activity for fulfilling local and more general demands on a relatively small available volume of blood.

The factors concerned with the local and general shifts or changes in volume of blood are not entirely known. They are related to such factors as sympathetic activity, digestion, sleep, psychic phenomena and many others. They are, probably, mainly concerned with thermal regulation and tissue nutrition, growth and repair. Although the changes of blood volume in tissues may not appear to be highly organized or orderly, they are, most probably, not just fortuitous phenomena.

Dr. Landis:

Have any chemical analyses been made on the fluid coming into the tissues, during prolonged standing or venous obstruction?

Dr. McMaster:

The amounts of fluid permitted to flow from the tissues into the injecting apparatus were too minute to employ for such purposes, and further, during the tests, the fluid in the tissues became mixed with unknown quantities of the test fluid initially introduced into the skin in order to study the pressure conditions there.

Dr. Mudd:

Studies have recently been made of the relative efficiency of intradermal, subcutaneous, and intravenous injections of various antigens in calling forth antibody formation. The findings presented here show that antigen, injected intradermally, is introduced, to a certain extent, directly into the lymphatics. As a result, some antigen is brought to the regional lymph nodes in high concentration, thereby affording a greater stimulus to antibody formation.

Plate 1

Figures 1a, 1b, and 1c (Magnification, x4). Successive stages in the distribution of dye introduced into the skin of the volar surface of the normal forearm. The photographs were selected from a moving picture film to show the course of events.

Figure 1a illustrates the entrance of dye into the rich network of lymphatic capillaries lying in the subpapillary layer of the corium. The photograph was taken 30 seconds after dye introduction began.

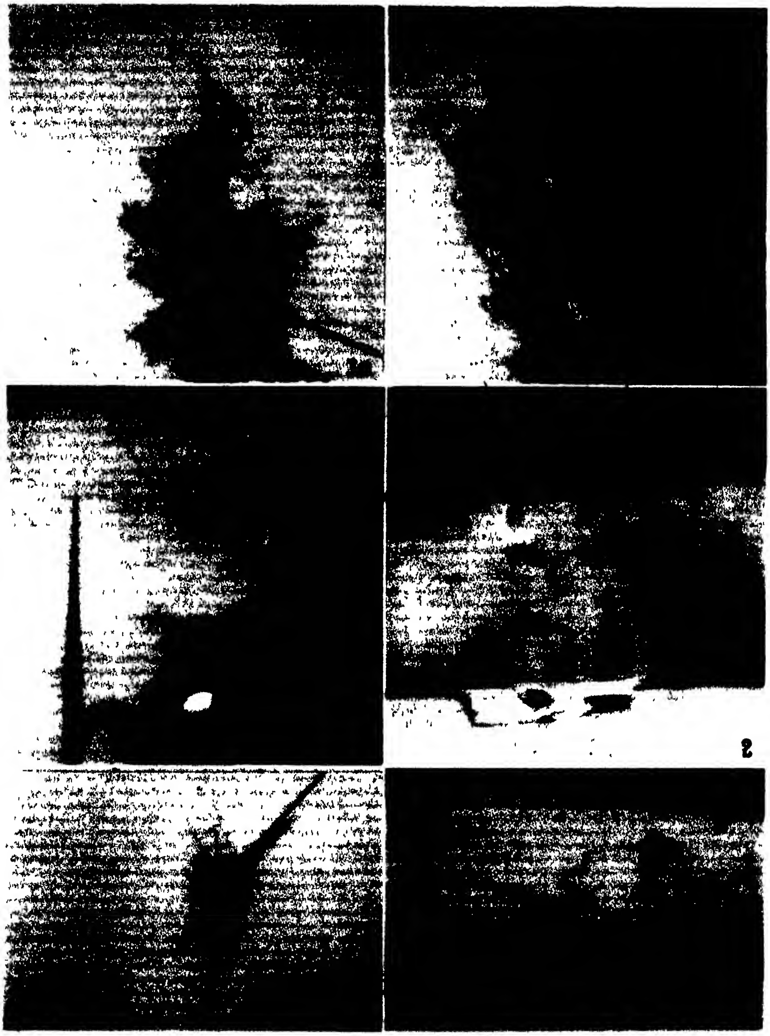
Figure 1b, taken about 5 minutes later, shows the spread which had occurred after the needle was removed.

Figure 1c shows the edge of the network of dye-containing lymphatic capillaries, half an hour after the introduction of the dye, and a pointer indicates a deep subcutaneous lymphatic draining away dye greatly diluted with lymph. The diluted dye seen through the skin looked like a blue streamer running up the arm.

Figure 2. A reduced photograph of the forearm of a normal subject, showing streamers arising from two injections of dye and beginning to extend up the arm.

______ 3a (Magnification, x1½). Dye introduced into the skin of the volar surface of the forearm, during a period of venous obstruction brought about by inflation of a pressure cuff on the upper arm. No streamers developed during the obstruction.

Figure 3b. Part of the intense streamer that had formed within 2 minutes after the onset of an intense reactive hyperemia which appeared a few seconds after the release of the obstruction.



PHILIP D. McMASTER: LYMPH FORMATION AND FLOW

PHILIP D. McMASTER LYMPH FORMATION AND FLOW

Plate 2

Figures 4a (Magnification, x1½) and 4b. The introduction of dye into the lymphatic capillaries of the normal ankle (figure 4a), and the short double streamer developing from this area as it appeared 20 minutes later (figure 4b).

Figure 5a (Magnification, x1½). The appearance of lymphatic capillaries in the skin of a swollen, edematous ankle of a patient suffering from cardiac insufficiency. The photograph is reproduced at the same magnification as figures 4a and 6a and was taken after the introduction of the same amount of dye in each instance. The lymphatic capillaries are widely dilated.

Figure 5b. A reduced photograph taken half an hour after the introduction of the dye. Although the patient was rapidly losing his edema at the time the test was made, there was no evidence of streamer formation.

Figure 6a (Magnification, x1½). Dilated lymphatic capillaries in the edematous skin of the ankle of a youthful patient suffering from nephrosis. The photograph was taken at the height of the edema.

Figure 6b Long and intense streamers, as they appeared 20–25 minutes after the introduction of the dye solution. The test was made during a period of spontaneous diuresis and the streamer formation was much more rapid and apparent than in previous tests made during periods when edema was increasing.

Figure 7 (Magnification, x3). Lymphatic capillaries around a healing punctate burn in the ear of an anesthetized mouse. The burn, made 9 days before the photograph was taken, had perforated the ear and rapid healing was in progress. A suspension of India ink in 5 per cent gelatin solution, introduced into the lymphatics of the tip of the ear, can be seen in twig-like, newly-formed lymphatics which had grown into the tissues that were closing the perforation. Proximal to the healing burn, there is an abnormally rich plexus of lymphatics situated in relatively healthy tissue from which active absorption seemed to be occurring, as judged by the rapid clearing of the contents of the vessels there.

Figure 8 (Magnification, x3½). Lymphatic participation in the repair of incisions. Dye introduced into the lymphatics at the tip of the ear of an anesthetized mouse, 5 hours after a transverse incision had been made in the skin of the upper surface. Colored fluid passed through the lymphatics and escaped from their severed ends into the incision. Open lymphatics proximal to the incision took up the colored fluid and drained it toward the base of the ear.

Figure 9 (Magnification, x17). Demonstration with India ink of the lymphatic plexus about a healing wound in the ear of the mouse. Nine days previously, the skin had been incised; healing was progressing well. On the left side of the healing incision, several reconstituted lymphatics are seen carrying ink directly through it.

THE SIGNIFICANCE OF LYMPHATIC BLOCKADE IN IMMUNITY

By Valy Menkin*

Department of Pathology, Duke University School of Medicine, Durham, North Carolina

I have been requested to review some of our earlier studies on lymphatic blockade. I shall start with the observations dealing with the fixation or retention of inert material and of viable organisms at the site of an acute inflammation. The literature on the subject has been extensively reviewed by the writer in a recent monograph, and, therefore, will be omitted in the present communication.[1]

Trypan blue in saline, in doses of 1.5 to 5 cc., injected into the normal hypodermis or the peritoneal cavity of a rabbit, diffuses readily to the tributary lymphatic vessels and nodes. When an acute inflammatory reaction is previously induced in one of these areas by the preliminary introduction of an irritant (such as aleuronat), the dye, subsequently injected, fails to diffuse readily to the tributary lymphatic nodes. With such highly irritating material as aleuronat, this local retention of the dye may be found to occur as early as 30 minutes after the injection of that irritant.[2] To illustrate the point, results of several such experiments are listed in TABLE 1.

Table 1

Retention of Trypan Blue at the Site of Subcutaneous Inflammation

Rabbit no.	Interval between injection of irritant (aleuronat and starch mixture) and of dye	Duration of inflammation	Presence of dye on normal side		Presence of dye on inflamed side	
			Lymph of efferent lymphatic	Lymph node	Lymph of efferent lymphatic	Lymph node
	hrs.:min.	hrs.:min.				
2	0:00	2:00	+	+	+	+
4	0:30	3:30	+	+	0	0
5	1:00	3:40	++	++	0	0
14	23:30	28:30	+	+	0	0
19	46:00	47:30	+++	+++	0	0

* Present address: Department of Surgical Research, Temple University Medical School, Philadelphia, Pennsylvania.

It is clear from such data that, not only with such intense irritants as aleuronat, retention of dye occurs extremely early. Furthermore, it is also evident that the dye is fixed quite effectively, for at least several hours, in such an acutely inflamed area of the skin (*cf.* Rabbit No. 14, in which the dye was held for at least as long as 5 hours in the inflamed area).

Such results have also been duplicated with ferric chloride, a foreign protein (horse serum), particulate matter (graphite) and microorganisms (*Bacillus prodigiosus*).[3, 4, 5]

Furthermore, it was also observed that these substances, as well as bacteria, when introduced into the circulation, rapidly accumulate into an acutely inflamed focus, provided the inflammatory reaction is not of too long duration. Capillary thrombosis tends to restrain the free seepage of such material into the extracapillary spaces in an inflamed area of long standing, *i.e.*, of 48 hours or over.

Fixation, or local retention of substances or of microorganisms at the site of an acute inflammation, is gauged by the inability to recover the material under study in either the tributary lymphatics or in the circulating blood. At the same time, it is important to determine that the material is actually held in the inflamed area. The following protocols with the use of *Bacillus prodigiosus* establish clearly these two points:

TABLE 2*

PRESENCE OF *Bacillus prodigiosus* IN RETROSTERNAL LYMPH NODES AFTER INTRAPERITONEAL INJECTION

Experiment	Interval between injection of irritant and of bacteria	Total duration of inflammation	Number of colonies recovered from retrosternal lymph nodes	
			After injection of bacteria into inflamed peritoneal cavity	After injection of bacteria into normal peritoneal cavity
	hrs.:min.	hrs.:min.		
1	4:10	6:00	6	150
2	15:05	21:10	3	45
3	15:30	21:40	2	38
4	22:45	25:45	0	Innumerable
5	26:18	29:50	39	Innumerable
6	24:45	40:10	7	Innumerable

* J. Exp. Med. 53: 652. 1931.

From the data of TABLE 2, it is evident that *Bacillus prodigiosus*, injected into an inflamed peritoneal cavity, fails to disseminate to the tributary lymphatic nodes in the retrosternal region as readily as under normal circumstances.

TABLE 3†

RETENTION OF *Bacillus prodigiosus* AT SITE OF SUBCUTANEOUS INFLAMMATION

Experiment	Interval between injection of irritant and of bacteria	Total duration of inflammation	Number of colonies recovered	
			Inflamed area	Normal area
	hrs.:min.	hrs.:min.		
1	1:30	3:30	175	37
2	4:00	6:00	225	175
3*	4:10	6:00	250	125
4	26:00	30:00	115	8
5	26:18	29:50	50	6

* The site of inflammation was located in the peritoneal cavity; see Experiment 1. Table II, for related data on retrosternal lymph nodes.
† J. Exp. Med. 53: 653. 1931.

The foregoing observations (TABLES 2 and 3) indicate that the inability of *Bacillus prodigiosus* to drain freely to the tributary lymphatics is, in all probability, due largely to the fixation of the bacilli *in situ*, at the area of acute inflammation.[5]

Similar observations have also been made with a metallic salt (ferric chloride). Incidentally, it is also conceivable that the accumulation of a metal from the circulation into an inflamed area, and its subsequent fixation in such an area, may alter the ultimate course of development of the inflammatory reaction. Furthermore, it is also possible that this type of localization from the blood stream may prove useful in roentgenological studies, provided the correct type of metal may be found which would, by accumulating in the inflamed area, augment its opacity.

In connection with the effect of an accumulation of a metallic salt in an inflamed area, it is noteworthy to recall that repeated intravenous injections of 0.25 per cent of ferric chloride in tuberculous rabbits are followed by an accumulation of iron in the tuberculous foci of the lung.[1] Concomitantly with this accumulation, the course of development of tuberculosis is retarded and the animals, apparently, survive longer than the controls. The clinical application of these observations is

further suggested by the ability of tuberculous patients to tolerate intravenous injections of .0625 per cent of ferric chloride,[6] and even up to 0.25 per cent concentration of the salt in saline.

The fixation of a foreign protein at the site of an acute inflammation may also be of value in our further understanding of the role of antibodies in specific inflammatory processes.[4] It is conceivable that their accumulation and fixation in an area of injury may help to explain the significance of such antibodies in furthering the localization of antigenic substances at the site of specific inflammation, as indicated by the studies of Rich and of Cannon.[7, 8]

In brief, fixation, by occurring at a very early period in the development of an acute inflammation, allows an interval of time for the relatively sluggish leukocytes to assemble for the purpose of phagocytosis. Without this immunological reaction, systemic infections, with inflammation, would be the rule rather than the exception.

THE MECHANISM OF FIXATION WITH INFLAMMATION

Fixation has been shown, in numerous early studies, to be referable to the presence of a fibrinous network in tissue distended with edema, and to the occlusion of the tributary lymphatic vessels at the site of inflammation.[9] The ultimate picture is an interference with the normal lymphatic drainage from the inflamed area. The result is essentially a lymphatic blockade. Both fibrinous deposits and thrombosis of lymphatics may be involved in the mechanism of fixation, or else only one of these two factors may predominate. For instance, in skin infections of rabbits induced by *Pneumococcus* Type I, occlusion of lymphatics seems to be of paramount importance, with a relative scarcity of fibrin deposits in the injured tissue spaces.[10] On the contrary, in burned areas, the lymphatic channels show relatively fewer occluding thrombi in the form of fibrinous plugs, but a network of fibrin in the burned tissue may form a conspicuous feature.[11] Other factors, under special circumstances, may reinforce the mechanism of fixation. These may involve the presence of antibodies in specific types of inflammatory reactions, precipitation or flocculation, and, perhaps, adsorption on fibrinous strands.[1] However, it is to be stressed that lymphatic blockade seems to be the primary basic mechanism involved in the various forms of non-specific inflammation. The local inflammatory edema is largely referable to the inability of occluded lymphatics to cope with the excess of fluid which, in turn, seeps both from the circulation, owing to a local increase in capillary permeability, and probably also

from cells, due to an increase in cellular permeability at the site of inflammation. Further information is desirable on the exact relation of lymph flow to increased capillary permeability, as the inflammatory reaction progresses. The observations on the mechanism of fixation have been further substantiated, to a large extent, by Drinker and his collaborators, in burned skin areas of dogs, and by Lurie, in his studies on the mechanism of the Koch Phenomenon in tuberculous guinea pigs.[12, 13] Glenn, Peterson, and Drinker demonstrated that lymphatic blockade can be readily accelerated in burned areas by the direct introduction of tissue extracts, which, in turn, hasten local coagulation.[12] Lurie showed that, in the reinfected guinea pig, the lymphatics adjoining the site of reinfection become thrombosed.[13] Furthermore, the fibrinous network in the inflamed areas of the guinea pig forms a closely knit, sieve-like structure which effectively prevents the dissemination of the bacilli.[13]

The evidence, indicating that the mechanism of fixation in an inflamed area is primarily referable to obstruction of free lymphatic drainage, has been frequently discussed in the past and will, therefore, not be reiterated here in detail. The reader is referred to some of these past publications.[1, 9, 14] The pertaining observations are merely mentioned, as follows:

1. Microscopic examination at the site of an acute inflammation reveals the presence, in varying degree, of a fibrinous network and of thrombosed lymphatics.

2. If various materials or bacteria are incapable of disseminating from a focus of severe inflammation, owing to mechanical obstruction in the form of lymphatic blockade, for the same reason, similar substances injected at the periphery of the area should fail to enter it. This is precisely what happens, as shown by observations with trypan blue and bacteria. Burrows has confirmed these experiments with India ink (1932).[15]

3. Urea, as a protein solvent, in large concentration, has been found to favor the solution of fibrin. The introduction of 50 per cent urea along with an irritant, such as aleuronat, tends to inhibit fixation by preventing the formation of fibrinous plugs, particularly in the lymphatics.[14]

In brief, three types of unrelated evidence have been advanced as follows: (1) morphological, (2) experimental (*i.e.*, by the injection of material at the periphery of an acutely inflamed focus and a study of its inability to penetrate into the area) and, finally, (3) the chemical

observations with urea which dissolves fibrin. All these three types of observations have led to the same conclusion in regard to the basic mechanism of fixation in acute inflammation, namely, that of lymphatic blockade. As mentioned above, other secondary factors may reinforce the basic mechanism, under special sets of circumstances.

THE MECHANISM OF INVASIVENESS OF PYOGENIC BACTERIA

The rapidity and the intensity with which an inflamed area becomes "walled off" by lymphatic blockade or coagulated plasma, or by both, is an index of the capacity of the irritating material to disseminate from the particular focus of acute inflammation which it has induced. In other words, the degree of tissue injury caused by a microorganism serves as a criterion, at least to some extent, of its own invasiveness. In this way, the invasiveness of a microorganism can be considered to bear an inverse relationship to the degree of cellular injury which it itself engenders at the point of entry. It is well known that *Staphylococci* caused a localized type of lesion, in contrast to the high invasive capacity of the hemolytic *Streptococci*. Observations have indicated that the difference in the invasive ability of some of these pyogenic organisms is, in large part, referable to the rapidity and the effectiveness with which they are capable of blocking adequate lymphatic drainage.[10] To illustrate, several experiments with three different pyogenic microorganisms are listed in TABLE 4.

TABLE 4

Microorganism injected	Time of establishment of lymphatic blockade as indicated by the inability of trypan blue to diffuse to the regional lymphatics						
	hours : minutes						
Staphylococcus aureus	0	0:30	1:03	1:43	3:43		21:50
	++	+++	faint trace	0	0		0
Pneumococcus Type I	1:48	4:17	6:06	17:38	20:45		
	+	trace	trace	0	0		
Streptococcus hemolyticus	1:44	6:07	17:05	17:50	23:45	30:03	48:16
	+++	+	+	++	++ to +++	+	0

It is quite clear from an examination of TABLE 4 that *Staphylococcus aureus* induces a sufficiently injurious reaction in skin tissue, so that, after about one hour, free drainage of dye to the efferent lymphatic vessels is hampered. *Pneumococcus* Type I occupies an intermediary position, whereas hemolytic *Streptococci* induce such a mild initial local injury that they take almost two days finally to block the lymphatic channels. These observations have been fully correlated with the histological appearances of the draining lymphatics. The inability of the dye to diffuse from the site of inflammation coincides with the histological establishment of lymphatic blockade. These observations present an interesting paradox. The most feared organisms are the ones that cause the least degree of injury at their point of entry. These, therefore, display great invasive ability through patent lymphatics, whereas microorganisms that are highly necrotizing correspondingly tend to be localized *in situ,* owing to an early development of lymphatic blockade. The patency of the lymphatic vessels determines, in large part, the invasive properties of a microorganism.[16]

In the case of *Staphylococcus aureus,* the additional exotoxic material produced by this organism enhances the localizing effect, for the broth filtrate of cultures of this organism is itself capable of causing prompt lymphatic blockade, *i.e.,* besides the organism *per se.*[17] Such is not the case with either the filtrate of *Pneumococci* Type I or that of the strain of hemolytic *Streptococci* used. It is conceivable that this potent exotoxin elaborated by the *Staphylococcus aureus* accounts in large measure for the severe necrotizing effect. There is some evidence that this injurious toxic material is somewhat similar, if not identical, to leucocidin. The absence of such powerful exotoxin in the filtrate of the strain of hemolytic *Streptococci* utilized may, in part, explain the relatively mild initial reaction at the point of inoculation, caused by these particular microorganisms. The fibrinolytic property of *Streptococci* scarcely explains the difference in invasiveness between *Staphylococci* and *Streptococci,*[18] for the simple reason that, under normal circumstances, in the rabbit (on which all our observations were made), the clot is not dissolved by the fibrinolysin in question.[10] It is more likely that, if the fibrinolytic factor enters at all in *in vivo* human streptococcal infection, this factor would reinforce the basic mechanism involved, namely, the further opening of the lymphatic channels. In the same way, Walston and the writer showed that "staphylocoagulase," responsible for the clotting principle in *Staphylococci,* is not primarily responsible for the localizing effect of the microorganism.[11]

The prompt mechanical obstruction to lymph flow by *Staphylococci* seems primarily referable to the powerfully necrotizing action *per se* of the microorganisms and of its soluble toxin. This conclusion was reached by effectively dissociating "staphylocoagulase" production from the factor responsible for local fixation.

THE PROBLEM OF INVASIVENESS AND VIRULENCE

The foregoing observations have led to an investigation of the basic relationship involved in the question of invasiveness and virulence. The two terms have often in the past been erroneously confused and used interchangeably. *Virulence* refers to the innate toxic property of a microorganism, whereas *invasiveness* refers to its disseminating capacity from the point of entry. The two properties can be studied more or less separately. For instance, the invasiveness of a microorganism can be interfered with by superimposing in the area of an inoculated infectious agent a severe inflammatory irritant which *per se* induces prompt lymphatic blockade. The dissemination of the original microbe is thereby delayed, and, in certain cases, may actually be ultimately disposed, in large part, at the site of inoculation, by the developing inflammation. In other instances, a retardation of the spread of bacteria may merely delay their systemic invasion, which, in turn, is followed by an increase in the longevity of the host. Such experiments were performed on rabbits, the skin of which was infected with a strain of *Pneumococcus* Type III. The immediate superimposed injection of aleuronat, turpentine, or *Staphylococcus aureus* at the same site accelerated the obstruction of lymphatics, delayed the invasiveness of the *Pneumococci*, and finally increased the survival time of the infected animal. The virulence of the *Pneumococci*, however, remained unaltered. Culturing these microorganisms, from the original site of cutaneous inoculation, or from the blood at post mortem, indicated that the original virulence had in no way been reduced. This could be readily demonstrated by studying the effect on normal rabbits subsequently inoculated with microorganisms recovered from such sites.[20]

In brief, these various studies clearly indicate that invasiveness is regulated by the local patency of lymphatics—hence, the significance of these channels which drain an area of acute inflammation. From this point of view, inflammation plays an important role in immunity as the regulator of bacterial invasiveness. The invasive capacity of a microorganism may be considered to be inversely related to the extent

of local injury which it induces at its point of entry. This may be conveniently expressed by the formulation $D = \frac{Kt}{I}$, where D refers to dissemination, and I, to the intensity of local injury. It is also clear that the dissemination is probably a direct function of time (t). K may be conveniently included as a constant referring to the type of irritant involved and, perhaps, to the anatomical location of the acute inflammation.[1]

THE ROLE OF LEUKOTAXINE AND OF NECROSIN IN THE MECHANISM OF LYMPHATIC OBSTRUCTION

Before concluding this discussion, it might be well to discuss briefly some of the chemical factors probably involved in the ultimate mechanism of lymphatic blockade.

One of the initial reactions in inflammation is an increased capillary permeability. This is, perhaps, true also of the lymphatics. The various studies on mild or initial tissue injury in which the lymph flow is increased suggest that this may be a possibility.[11, 12, 21] The factor primarily concerned with the increased capillary permeability in inflammation is leukotaxine, apparently a relatively simple polypeptide to which there seems to be attached a prosthetic group, the nature of which is as yet unknown.[22] It is possible that the release of this substance may also enhance the permeability of the lymphatic vessels at the site of an acute inflammation. At any rate, the increased capillary permeability, largely referable to the presence of leukotaxine, allows the passage of albumin, globulin, and fibrinogen molecules from the circulation into the extracapillary areas. The presence of fibrinogen in contact with products of injured tissue doubtless favors the formation of a coagulum. The lymphatics are apparently more delicate in structure than the capillaries. They, in turn, become plugged with fibrin and a state of thrombosed or occluded lymphatic is then in evidence. In due time, the capillaries may likewise reveal the presence of thrombi, thus tending to shunt the area of local injury from the rest of the organism. This segregation of an injured part involves a new local and impaired circulation, change in pH, and a characteristic metabolism.

The pattern of injury in acute inflammation has been shown, in earlier studies, to be primarily referable to the euglobulin fraction of exudative material.[23] This factor can be readily dissociated, in exudates, from the fever-inducing substance, which, in turn, is apparently a glycopeptide (in view of recent observations by Dr. Frederick Bern-

heim and the writer). This pyrogenic factor has been termed *pyrexin.*[30] The injury factor is a true euglobulin or, at least, is associated with that fraction of exudates. Recent studies indicate that this substance, named *necrosin,* recovered in the euglobulin fraction of exudates, is frequently found in exudates which have an acid pH.[24] The earlier studies of the writer, recently confirmed by Rugiero and Tanturi in Argentina,[25] have demonstrated that, with the progress in intensity of an acute inflammation, a state of local acidosis develops.[26, 27] It is this rise in the hydrogen-ion concentration that conditions the cytological picture at the site of inflammation. Below a pH of 7.0, the polymorphonuclears appear markedly damaged and the macrophages are normal. When the pH falls below 6.5, all types of leukocytes show signs of severe damage and, virtually, a state of suppuration ensues. Pus formation is essentially a function of the pH. Now, it is quite clear from these recent studies that, as the exudate increases in acidity, a corresponding rise in necrosin formation tends to occur. This substance, which contains a proteolytic enzymatic activity, is highly injurious to tissue. It induces, among other changes, thrombosis of lymphatics (PLATES 3 and 4) and of small vascular channels. Consequently, the liberation of necrosin is probably of significance in inducing further lymphatic blockade, as the inflammatory reaction progresses in intensity and develops a local acidosis.

Necrosin is absent in normal blood serum, but it is often recovered from the serum of an animal with a concomitant acute inflammation. This fact may be of significance in furthering our understanding of the role of foci of infection on organs located at a distance. In this connection, it has been shown that necrosin induces some sort of hepatic injury. Repeated intravascular injections of this substance produce, as a rule, various types of liver cell injury. This may consist of granular debris, vacuolation, fat deposits, or abundance of glycogen, even though the animals have been starved for about a day prior to examining the organ. It is conceivable that the glycogen is perhaps referable to cellular injury in the liver, with consequent deamination and the formation of carbohydrate products.[28, 29] The kidney, particularly, frequently shows evidence of damage, though not exclusively, in the lining epithelial cells of the tubules. These studies are being further pursued.

BIBLIOGRAPHY

1. **Menkin, V.**
   1940. Dynamics of Inflammation. MacMillan Co. New York.
2. **Menkin, V.**
   1929. J. Exp. Med. 50: 171.

3. **Menkin, V.**
1930. J. Exp. Med. **51**: 879.
4. **Menkin, V.**
1930. J. Exp. Med. **52**: 201.
5. **Menkin, V.**
1931. J. Exp. Med. **53**: 647.
6. **Menkin, V.**
1937. Am. Rev. Tuberc. **35**: 134.
7. **Rich, A. R.**
1933. Bull. Johns Hopkins Hosp. **52**: 203.
8. **Cannon, P. R., & G. A. Pacheco**
1930. Am. J. Path. **6**: 749.
9. **Menkin, V.**
1931. J. Exp. Med. **53**: 171.
10. **Menkin, V.**
1933. J. Exp. Med. **57**: 977.
11. **Menkin, V.**
1933. Proc. Soc. Exp. Biol. & Med. **30**: 1069.
12. **Glenn, W. W. L., D. K. Peterson, & C. K. Drinker**
1942. Surgery **12**: 685.
13. **Lurie, M. B.**
1939. J. Exp. Med. **69**: 555.
14. **Menkin, V.**
1932. J. Exp. Med. **56**: 157.
15. **Burrows, H.**
1932. Some Factors in the Localisation of Disease in the Body Baillière, Tindall, & Cox. London.
16. **Menkin, V.**
1938. Physiol. Rev. **18**: 366.
17. **Menkin, V.**
1935. Am. J. Med. Sci. **190**: 583.
18. **Tillett, W. S., & R. L. Garner**
1933. J. Exp. Med. **58**: 488.
19. **Menkin, V., & H. D. Walson**
1935. Proc. Soc. Exp. Biol & Med. **32**: 1259.
20. **Menkin, V.**
1936. J. Infect. Dis. **58**: 81.
21. **Hudack, S. S., & P. D. McMaster**
1933. J. Exp. Med. **57**: 751.
22. **Menkin, V.**
1938. J. Exp. Med. **67**: 129.
23. **Menkin, V.**
1943. Arch. Path. **36**: 269.
24. **Menkin, V.**
1945. Fed. Proc. **4**(1).
25. **Rugiero, H. R., & C. A. Tanturi**
1942. La Semana Medica. **8, 13.**
26. **Menkin, V.**
1934. Am. J. Path. **10**: 193.

27. **Menkin, V., & C. R. Warner**
1937. Am. J. Path. **13**: 25.
28. **Menkin, V.**
1941. Am. J. Physiol. **134**: 517.
29. **Menkin, V.**
1943. Am. J. Physiol. **138**: 396.
30. **Menkin, V.**
1945. Arch. Path. **39**: 28.

DISCUSSION OF THE PAPER

Dr. Eugene Opie (*Rockefeller Institute for Medical Research, New York, N. Y.*):

The site of inflammation in mammals is, with few exceptions, vascularized connective tissue at the site of injury, together with regional lymphatic vessels and lymphatic nodes. Carbon particles, bacteria, even those so large as anthrax bacilli or foreign red corpuscles, injected into the subcutaneous tissue, penetrate almost immediately to the sinuses of the regional lymphatic node and may be discovered there within several minutes after their introduction. Lymph vessels and lymphatic nodes undergo changes similar to those that occur at the site of entry. Local fixation soon retards the further dissemination of the inflammatory irritant.

Flow of lymph is accelerated by acute inflammation, and not retarded, as might be expected if lymphatic vessels were occluded by fibrinous coagula. This relation is well shown by the old experiments on lymph flow with inflammation cited by Cohnheim and by the recent observations of Drinker and his co-workers. Fixation of particulate material is doubtless brought about, in large part, by the fibrinous network that is formed within tissue spaces at the site of inflammation. If bacteria, *e.g.*, *Staphylococcus aureus* or carbon particles of India ink are added to oxalated plasma, the fibrinous clot that is formed on addition of calcium assembles these particles, so that the fluid surrounding the shrunken clot has become quite clear. Formation of a fibrinous coagulum at the site of inflammation mechanically fixes particulate matter and, though similar fixation may occur in lymphatic channels and in the sinuses of lymphatic nodes, obstruction of lymphatic channels has little, if any, part in the fixation of bacteria or other particulate matter.

Dr. Max Lurie (*Henry Phipps Institute, Philadelphia, Pa.*):

First, I should like to corroborate Dr. Opie's observation on the capacity of the clotting process to remove particles suspended in the plasma. If to citrated plasma, containing uniformly suspended, living tubercle bacilli, is added enough calcium to form a firm clot, 90 per cent of the tubercle bacilli will be found by culture in the clotted fibrin, whereas, in the supernatant serum, very few bacilli remain.

That the intensity of the injury exerted by the inflammatory agent on the tissues is an important factor in the fixation of that agent at the portal of entry, can be readily demonstrated in experimental tuberculosis of the rabbit and guinea pig.

Tubercle bacilli suspended in melted agar, containing trypan blue, were injected subcutaneously in normal and immunized rabbits. Both the tubercle bacilli and the trypan blue were flushed to the draining lymph nodes more rapidly in the immunized than in the normal animal, and the sinuses of the lymph nodes draining the site of inoculation in the immunized rabbit were crowded with polymorphonuclear leucocytes, whereas these sinuses in the normal animal were free of exudate cells. The flow of lymph from the site of reinfection is greater than that from the site of primary infection.

Obviously, the increased lymph flow from the site of intensified inflammation of reinfection had swept these entities, as well as agar particles, to the lymph nodes more quickly than did the lymph flow from the site of the less intense inflammation in the normal animal. Nevertheless, it may be said parenthetically, the bacilli which reach the lymph nodes in the immunized animal so quickly are

markedly suppressed in their multiplication, whereas the bacilli draining the site of infection in the normal animal, though they arrive at the lymph nodes more tardily, find conditions suitable for their rapid growth, so that, within two weeks after inoculation, they are a thousand times more numerous in the latter than in the former lymph nodes.

If similar experiments are performed on immunized and normal guinea pigs, it is found, on the contrary, that both the bacilli and the trypan blue are retarded in their dissemination to the draining lymph nodes in the immunized, as compared with that in the normal, animal. Clearly, therefore, the passage of particles from the site of reinfection in a guinea pig is more effectively impeded than from a similar site in the rabbit.

On comparing the character of the inflammatory process at the site of reinfection, in these two species, the explanation of the observed difference becomes clear. In the guinea pig, there is a dense fibrinous clot at the site of reinoculation, and the draining lymphatics are thrombosed by a fibrinous network. In the rabbit, however, there is a loose, wide-mesh, fibrinous deposit at the site of reinfection, and the draining lymphatics are widely open, containing but a few shreds of fibrin.

Now, it is well known that the tubercle bacillus and its products exert much greater toxic effects on the tissues of the immunized guinea pig than on those of the sensitized rabbit. Clearly, therefore, the destructive effect of the tubercle bacillus on the tissues of the immunized guinea pig releases enough thrombokinase to produce a dense clot at the site of reinfection, which entraps the bacilli in the process, as stressed by Dr. Opie. Furthermore, the same thrombokinase, released in the guinea pig, is sufficient to clot the lymph in the lymphatics, thus reinforcing the blockade around the site of reinfection. In the immunized rabbit, however, due to the lesser sensitivity of its tissues to the tubercle bacillus, less thrombokinase is released, the clot at the site of reinfection is less dense, and the draining lymphatics remain open. Hence, due to the increased lymph pressure at the site of reinfection, as compared with that at the site of primary infection, in a rabbit, the lymph flow is greater in the former, and the bacilli are swept onward through the open lymphatics more rapidly than in the normal rabbit.

Dr. Herbert L. Davis (*Ethicon Suture Laboratories, Division of Johnson & Johnson, New Brunswick, N. J.*):

This splendid work of Dr. Menkin has been followed with much interest. Some of the phenomena described may possibly be interpreted in terms of the concepts of the colloid chemistry of such systems, and the application of colloid principles may guide further progress in this most promising field.[1]

In dealing with living tissues, one has in hand materials which the colloid chemist characterizes as hydrophilic systems, which means that there exists a large affinity between the solid components and water. Closely allied to the degree of hydration shown by such solids, is their electrical charge, so that their dispersibility is a result of the interplay of both these forces. Thus, gelatin is dispersible by water alone, albumen has a high degree of hydration, but requires some charge to disperse it, casein has little affinity for water and must be highly charged to disperse, while collagen is not dispersible until it is broken chemically into smaller units passing toward the gelatin stage. Precipitation of proteins from sols by salts such as ammonium sulfate involves dehydration by high salt concentration, and removal of charge by bringing the proteins near to their isoelectric point.

Living tissue is characterized, not only by its hydration and charge, but also by the fact that its structure is comparable to that of a gel, in which the solid phases and the liquid phase are both continuous. This means that, in the brush-heap-like structure, a particle small enough might pass from any point in either phase

[1] For the point of view presented herein, cf:

Alexander, Jerome. Colloid Chemistry 5. Theory and Methods, Biology and Medicine. Reinhold Publishing Corp., New York. 1944.

Hartman, R. J. Colloid Chemistry. Houghton Mifflin Company, New York. 1939.

Gortner, R. A. Outlines of Biochemistry. John Wiley & Sons, Inc., New York. 1938.

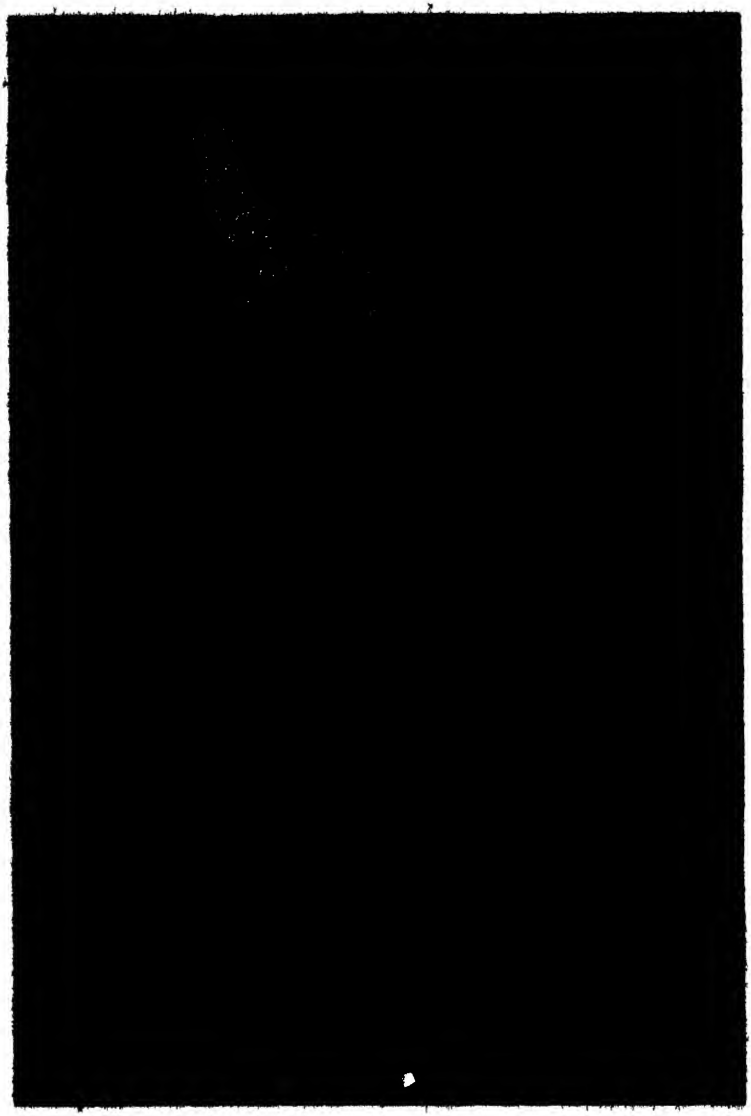
to any other point in that phase, without leaving that phase. At selected points or areas, the structural elements of the solid phase may be so concentrated, and in such a state of hydration or swelling, as to preclude any free passage, even of water itself (cell walls). At other points, there may be free passage of proteins, cells, and foreign particles.

If living tissues have such a structure, this structure may then be modified by chemical additions. One important factor may be the alteration in the degree of hydration. If a gelatin gel be immersed in pure water or a dilute salt solution, it will swell, while a concentrated salt solution will cause marked contraction. While this is related to osmotic pressure, the Hofmeister or lyotropic series arranges anions and cations in terms of their swelling power on such gels. Thus, if isotonic solutions of sodium salts be used, sulfate, tartrate, and citrate cause contraction, while thiocyanate, iodide, bromide, and chloride cause swelling of the gel. A third factor is that shown by the acids and alkalis which, in both cases, swell the gel, but at some intermediate point (the isoelectric point) the volume of the gel is a minimum. In all such systems, the electric charge of the solid phase is important; if this be negative (as is usually the case in tissues), then positively charged ions, colloidal particles, or cells would be readily absorbed, with consequent closing of the pores and reduction of permeability.

It may be that, in these findings of Dr. Menkin, we are dealing with similar phenomena. While, obviously, bacteria, cells, and protein aggregates do carry electrical charges and may mutually precipitate each other, as in precipitin reactions, it seem more probable that it is their metabolic products which exert an agglomerating or dispersing effect on the invisible, highly hydrated, structural elements, and, in this contraction or swelling, close or open up channels for diffusion of dye or organisms. Thus, aleuronat, ferric chloride, graphite particles, horse serum, *Bacillus prodigiosus*, and *Staphylococcus aureus* (or its broth filtrate) show coagulating effects; *Pneumococcus* Type I seems to have little effect; while urea and hemolytic *streptococci* exert a swelling effect.

It is quite probable, as Dr. Menkin has pointed out, that many such mechanisms play a part in determining the diffusion of a dye or the invasiveness of a microorganism. Extending these ideas to other polyvalent cations, coagulating anions or nonelectrolytes (including some of the anesthetics and wetting agents) may provide still more effective ways of limiting the invasion of pathogenic organisms. Conversely, such materials as urea may spread markedly the entrance of antitoxins, antibodies, or other remedial agents. It is a project fraught with large promise, and Dr. Menkin is to be heartily congratulated on the great progress made.

PLATES 3–4



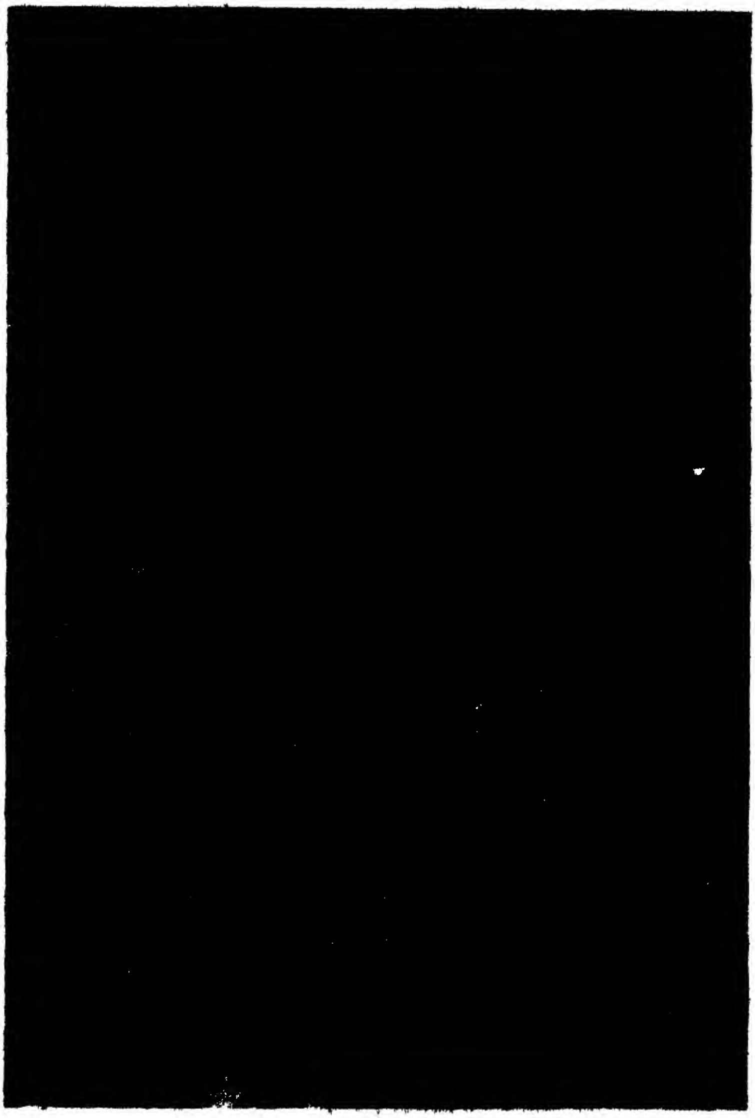
VALY MENKIN: LYMPHATIC BLOCKADE IN IMMUNITY

PLATE 3

Lymphatic vessel plugged with fibrin in an area of acute injury in the skin of a rabbit, induced by the injection of necrosin derived from an acid exudate. The inflammation is one day old. (x 445)

PLATE 4

A lymphatic vessel occluded by a dense leukocytic thrombus in an area previously injected with necrosin The material was injected about one day prior (x 485)



VALY MENKIN: LYMPHATIC BLOCKADE IN IMMUNITY

EXTRAVASCULAR PROTEIN AND THE LYMPHATIC SYSTEM

By Cecil K. Drinker

Department of Physiology, Harvard School of Public Health, Boston, Mass.

In this paper, I return to my first interest in the lymphatic system and the issues which have intrigued me most since entering this field. In 1927, it was my good fortune to be working for Professor Krogh in his modest laboratory upon a residential street near the center of Copenhagen. There he had conceived and carried out the sagacious experiments which laid the foundations for our present conceptions of the capillary circulation. At that time, Krogh was checking and improving the methods he had employed to measure the colloid osmotic pressure of extremely small amounts of blood serum, a very essential technique for his laboratory, where most of the experimental work was being carried out upon frogs, often small frogs with little blood. The measurements of colloid osmotic pressure of blood serum verified the luckily correct findings of Starling[1] in 1896. The result of Krogh's labors and that of his pupils established, beyond question, the balance between capillary blood pressure and the colloid osmotic pressure of the blood proteins for rapid movements of water out of and into the blood.

In 1927, so far as I can recall our beliefs in these matters, it was the general conviction that the walls of typical blood capillaries, throughout the body, did not permit the escape, except some very small traces, of the blood proteins. It was agreed that the capillaries of the liver and the intestines allowed leakage of these compounds, but these were considered special cases, and the capillaries in most parts of the body were considered practically impermeable to the blood proteins. It is appropriate to point out that quantitative analysis of the protein in liver and intestinal lymph was the means of measuring this quite specific leakage from capillaries in two large areas.

In the early spring of 1927, Professor Krogh made an expedition on his bicycle and returned with a sack full of frogs. These, put in the small pond behind the laboratory, turned immediately and enthusiastically toward the task of producing more frogs. Many of them became quite edematous as this period passed, and in the laboratory

we speculated widely upon the cause of the edema. At the moment, some ductless gland derangement was most intriguing, but I am inclined to believe, today, that the frogs, having had nothing to eat through several dark and chilly Danish months, found themselves impelled into hard physical work by their domestic responsibilities, when, unfortunately, their blood proteins had been greatly lowered by a fairly long period without food. Whatever may have been the explanation, whether glandular or relatively simple, the laboratory had a large supply of edematous frogs. The edema fluid was mainly in the lymph sacs of the animals and was moving through the conventional lymph pathway back to the blood. This fluid or lymph was easy to collect, and, in a small series of frogs, contained 1.0 per cent of protein, and clotted on standing.[2] Since the normal concentration of protein in frog blood is about 3.5 per cent, this figure for extravascular protein was very high. Whether the lymph was collected from under the skin, in the hind feet, or from the abdominal cavity, the protein concentrations in the frog under examination were uniform.

It was impossible to contend that this extravascular protein had not come from the blood through the capillary walls. However, this simple means of accounting for it did not fit current belief, which found it inconvenient to admit that any such quantity of blood protein could leave the blood, even in the frog with four lymph hearts steadily engaged in pumping extravascular fluid back into the blood. We blamed the high lymph or tissue fluid protein upon some abnormality in the frogs, a physico-chemical expression of the Danish springtime.

In those days—and they are not so long past—our interest in the functions of the blood proteins, except for the problem of blood coagulation, was just beginning. Starling's theory of the blood proteins (that they accomplished balance between intravascular and extravascular water) was confirmed by Krogh and his pupils, and this function of the blood proteins covered their entire significance for a number of years. It did not occur to investigators that the normal proteins of the blood could prove an adequate source of nitrogen for the body. The walls of typical blood capillaries were thought to be very similar to the membranes used in the laboratory for measuring the osmotic pressure of serum. The tissue fluid was believed to contain only traces of protein, as evidenced by analyses of edema fluid collected from the subcutaneous tissue of patients with cardiac decompensation. Whatever might be the case in mammals, it appeared certain that, in frogs, the capillaries were quite permeable to the blood proteins. These

animals had an extravascular circulation motivated by four lymph hearts which returned the proteinized fluid bathing the body cells to the blood, thereby maintaining reasonable constancy of blood volume.

During the next few years, Ruth E. Conklin[3] investigated lymph formation and circulation in frogs. She found that, if the lymph hearts were stopped by curarization, there was a rapid loss of blood plasma into the lymph spaces, and that the plasma proteins could be washed out of the blood if the lymph hearts were stopped and Ringer's solution infused slowly through a vein.

By 1931,[4] it had become possible to shift our experiments to mammals, in the sense of comparing the protein concentration of cervical, hind leg, foreleg, thoracic duct, kidney lymph, and blood serum from dogs under barbiturate anesthesia. The problem which had arisen was the sorting out of several possible fluids originating from the blood. There was, first, the filtrate from the capillaries, then the tissue fluid, and last, the lymph. Of these extravascular fluids, the only one readily collected was lymph, and we had the temerity to suggest that the lymph might represent the composition of the tissue fluid in the area drained by the cannulated lymphatics. The conception of the situation, presented in this first paper on mammalian lymph and tissue fluid, seems to me, today, to have been better than a good guess, but it was severely deprecated in the next few years. What was actually written was: "One may consider that capillaries practically universally leak protein; that this protein does not reenter the blood vessels unless delivered by the lymphatic system; that the filtrate from the blood capillaries to the tissue spaces contains water, salts, and sugar in the concentrations found in blood, together with serum globulin, serum albumin, and fibrinogen in low concentration, lower probably than that of tissue fluid or lymph; that water and salts are reabsorbed by the blood vessels and the protein enters the lymphatics together with water and salts in the concentration existing in the tissue fluid at the moment of lymphatic entrance. The lymph from any given drainage area contains a varying amount of protein, dependent upon the amount of water absorption which has taken place in the region from which the collection is made, and represents a cross section of the tissue fluid of the area in question."

Nothing has caused me to feel this to have been a bad appraisal of the situation. As the years have passed, it seems to me that the evidence obtained through collection and actual analysis of tissue fluid and lymph indicates identity of lymph and tissue fluid rather than

difference. For example, Maurer[5] collected tissue fluid from the muscles of frogs, and found protein concentrations identical with those available for frog lymph, so that in this animal, in which, it is now admitted, blood capillaries are permeable to more than traces of protein, lymph and tissue fluid have the same composition.

What is, however, well enough for frogs, may be thought far from the mark in mammals. The quantitative evidence bearing upon the identity of composition of tissue fluid and lymph in mammals is fragmentary and, in the main, applies only indirectly to normal conditions. No one has ever collected tissue fluid from such a region as the subcutaneous region of normal mammals. The estimations of tissue fluid composition have been derived indirectly and stem from a period when the usefulness of extravascular blood protein was little understood. There are few actual figures for the composition of lymph and tissue fluid collected simultaneously from mammals. In dogs with lymphatic obstruction,[6] it was found that the protein content of edema fluid and lymph was identical. Such evidence, based upon abnormality of lymph drainage, may well be viewed skeptically. In mammals and birds, it is always possible to obtain a few cubic centimeters of pericardial fluid and, not infrequently, enough pleural and abdominal fluid to permit quantitative analysis of protein.[7] These fluids are derived from the blood. They differ in eventual position from ordinary tissue fluid, by being enclosed in endothelium-lined sacs. The protein content of these fluids is high. For pericardial fluid from 34 dogs, the average value was 1.7 per cent; from 7 rabbits, 2.16 per cent; from 4 monkeys, 1.71 per cent; from 2 cats, 2.42 per cent; from 1 rat, 2.07 per cent; from 2 hens, 3.53 per cent; from 2 ducks, 2.51 per cent. The average protein of the peritoneal fluid from 11 dogs was 2.61 per cent; and from 5 rabbits, 1.53 per cent. These figures are in the range of lymph protein concentrations gained from lymph leaving the same regions. I can see no reason to believe that the endothelium lining any of the great body cavities has absorptive, secretory, or other properties which might affect the composition of contained fluid. In my opinion, the high protein content of the fluids described is a characterization of tissue fluid in general.

One of the most distinctive characteristics of lymph capillaries is the readiness with which they are entered by all sorts of visible particles and by molecules which are unabsorbable by the blood vessels. It is this feature of the lymphatic apparatus which makes the system so important in connection with the spread of infections. However, the

most constant and important function of the lymphatics is, apparently, the unremitting removal from the tissues of excess blood protein. In all parts of the body, except the liver and intestines, the amount of lymph in movement toward the blood seems slight; and, while blockage of the lymphatics from a part may result in the slow development of edema and elephantiasis, the lymphatics, as a whole, have little place in the relief of edema. They may be filled with highly proteinized inflammatory exudates containing bacteria, dirt, grease, anything forced into the drainage area, but the actual volume of fluid moved is not great, except from the liver and intestines.

Viewing the lymphatic system through the mists of twenty years of unremitting work, I have slowly realized what is perhaps very obvious to those less preoccupied with details of function; namely, that, in mammals, the lymphatics have precisely the same functional task as in the frog. In that animal, one may recollect, we know that there is an extravascular circulation—a delivery back to the blood of lymph impelled by the four lymph hearts. In mammals, the propulsion of lymph through the valved vessels depends upon forces outside the system, such as the massaging effects of active and passive movements, but the accomplishments of the lymphatic system are quite the same as in the frog. They are the steady movement of extravascular protein back to the blood. This conception of an elaborate mechanism to deal with proteins which have left the blood vessels requires the acceptance of some degree of protein leakage from capillaries, everywhere in the body. If our concept of the function of the blood proteins had remained as it was in 1927 (solely that of the regulation of water movement between the capillary blood and the tissue), the mammalian lymphatics would seem to have little to accomplish, under normal conditions.

Years ago, there was evidence, to which physiologists paid no attention, that antibodies such as diphtheria antitoxin were probably globulins. Obviously, if such agents were to reach the ultimate sites of disease, they had to be distributed by the blood, and must leave the capillaries to neutralize disease processes in the tissues. Where inflammatory dilatation and abnormal permeability of blood vessels were present, it was easy to understand that substances of the molecular size of globulins would leave the blood stream to flood the diseased tissue. But evidence was presented which characterized the ability of antibodies to move out of the blood in a much more general way.[8] This evidence deals with antibody entrance into the lymphatics and

the failure to achieve a concentration capable of sterilizing the lymph, though temporary sterilization of the blood was readily accomplished.

The inadequately-formed concept of the blood proteins had, however, failed to progress, physiologically, beyond the supposition that they were a convenient agency for water-movement from blood to tissues and back again, until, in a brilliant and systematic series of researches, Whipple and his collaborators[9] showed that dogs could be fully supplied with nitrogen by intravenous administration of dog plasma. The idea that extravasated protein might supply nitrogenous compounds necessary for healthy, vigorous life gave the blood proteins a new physiological importance. It is not part of my task to summarize all the possible ramifications of the work from the Rochester Laboratory. It is, however, wholly appropriate to emphasize the degree to which they support the existence of a tissue fluid relatively rich in protein to supply the needs of surrounding cells. Whipple and Madden conclude the paper to which I have referred with this paragraph:

"Our concept of a large *protein pool* including the circulating plasma proteins and mobile cell proteins emerges from this discussion. The contributions to this pool derive largely from the liver, and the withdrawal from the pool may concern any body cell needing protein or capable of storing some surplus protein. From this protein pool may be derived hemoglobin, new plasma protein, or cell protein. The circulating plasma protein is the medium of exchange, and the body is solvent just so long as there is adequate protein supply for any emergency. When the body becomes insolvent, there may be a foreclosure due to disease, infection, or injury."

These are not simple imaginative excursions, but concepts which are based upon experimental evidence and the mental growth which emerges from years of battling in the troubled field of medical research It is my task to relate the widespread lymphatic system of mammals and the equally widespread distribution of the compounds we have called blood proteins (today, more aptly termed, body proteins).

No one today quarrels with the physiological concept of Claude Bernard that the environment of the body cells is held rigidly constant, and that the great variety of physiological reactions, dissected through so many years, has, as its invariable direction, the maintenance of this constancy. From this basic idea, another at once emerges. It is impossible to conceive constancy of chemical surroundings for living cells, which are using substances from the fluids in which they live and are giving off wastes to the same fluids, unless there is constant

change or renewal of the fluid. In the frog, the position of the lymphatic system for holding the extravascular liquid steady in composition is readily appreciated. If lymph return to the blood is prevented by stopping the lymph hearts, the frog cannot live, since he lacks the steady return of water, solutes, and, above all, blood proteins, to the blood vessels.

In mammals, notably man, it has been known for years that when lymphatic drainage of a part is interrupted, a brawny edema develops slowly, with a concomitant overgrowth of connective tissue and even epithelium. This cellular growth is not abnormal, as is the development of neoplastic tissue. It is infrequent that neoplasms occur in regions of lymphatic obstruction, though, with many other theoretical notions, blockade of lymph drainage has been blamed for the origin of new growths. What happens in the subcutaneous tissue of the leg of a dog[10] in which lymph drainage has been fairly effectively blocked is a relentless laying-down of fibrous tissue, so that the normal, insignificant, pliable layer of connective tissue overlying the muscles becomes several centimeters thick, and the part becomes permanently elephantiatic. The changes which occur are in the nature of what one might expect, if the part were mildly, generally, and continuously the site of chronic inflammation. While it is true that man and animals, with lymph blockage in a large tissue area, frequently experience erysipelatous attacks in the affected part, which intensify the development of elephantiasis, this does not mean that acute inflammation is essential for the production of the condition. The ultimate cause of elephantiasis is not known, nor have I introduced this discussion of lymphatic obstruction with any other idea than that of indicating how vital normal lymph drainage is to the mammal. The total amount of lymph present at any one time cannot be approximated, but lymph volume must vary, markedly and even fairly abruptly. Where the dimensions of lymph capillary areas have been measured and compared with similar figures for blood capillaries in the same tissue, the results are quite similar, so that, potentially, the lymphatics can prove a fairly large fluid reservoir. It is doubtful if they really function in this way, except in regions of inflammation, where the smallest lymphatics become widely dilated and can easily be shown to be filled with highly proteinized fluid, leucocytes, etc., derived from the blood. From most normal tissues—the legs, heart, lungs, head, and kidneys—of the dog, one can always collect small amounts of protein-containing lymph, providing appropriate measures are used to cause lymph flow.

The protein, including fibrinogen and prothrombin, is always present, but concentrations vary markedly from time to time. It appears that, in the mammal, there is a steady movement of protein out of the blood and into the tissue fluid, where part of it is utilized as a source of nitrogen by the body cells. Whipple's concept of a pool of available body protein treats of the lymphatic vessels as an agency for keeping this pool steadily in a state of change, making for reasonable constancy of cellular environment.

These considerations apply to the non-specialized parts of the body, as regards production and movement of lymph. By non-specialized parts, I mean practically all of the body, except the small intestine and liver. From these two regions, lymph flow is large, and the lymph collected is high in protein. Furthermore, the flow of lymph from the thoracic duct is considerable, except under conditions of starvation and water depletion. Under ordinary circumstances, it would seem that about 95 per cent of the lymph turned back into the blood through the thoracic duct is formed in the liver and intestines.

Recently, a fortunate clinical experience has given us better information upon the consequences of complete loss of thoracic duct lymph.[11] A 30-year-old colored woman was shot in the left side of the neck, one hour before admission to the hospital. The left internal jugular vein was ligated, two days after admission. During the operation, straw-colored fluid steadily welled up in the wound, so that the skin was not closed. After the operation, the dressings were rapidly saturated with this fluid, and, for the next six weeks, a ceaseless leakage of what was unquestionably thoracic duct lymph continued. The patient at once took the regular hospital diet, and, after she had eaten, the leaking fluid became milky. But she lost weight, at a rate of about five pounds a week, and her plasma protein fell to 3.5 per cent in just a month. A diet high in protein brought this to 4.6 per cent in thirteen days, but weight loss continued. Accordingly, in a second operation, the thoracic duct was ligated and the wound closed. For two weeks after this ligation, the patient had cramps after eating, but she gained 16 pounds in a month and three days, and was discharged, free from complaints.

In this healthy young woman, it was possible to observe the effects of practically complete loss of lymph entrance into the blood. This patient, while the thoracic duct fistula was open, apparently experienced only such return of lymph to the blood as was accomplished through the right lymphatic duct, which is concerned almost entirely with lymph flow from the heart and, particularly, from the lungs.

There are collateral vessels. These consist in connections between the thoracic duct and the right duct, and entrances of the thoracic duct lymph into veins at lower levels than the left subclavian vessel. In dogs, in which we know far more about these possibilities than in man,[12] it is apparent, not only from the reference given, but from our own experience, that potential drainage paths for thoracic duct lymph are available in most animals, but if the thoracic duct is wide open, so that no increased pressure factor is present to cause this lymph to shift flow into the right duct, then these collateral lines of lymph flow are not opened. In this patient, the important facts for us are that, with the lymph lost to the blood, the patient experienced a progressive loss of blood protein which apparently would have been fatal, had it not been interrupted by ligation of the thoracic duct. What this experience really teaches is that, if a healthy adult is deprived of normal delivery of extravascular protein to the blood by the lymph route, serious dislocation of body fluid composition begins to occur promptly. This significant human experiment has been attempted many times upon animals, but never successfully. I speak with some degree of feeling, which arises from a long effort to make a permanent thoracic duct fistula, by exteriorizing the subclavian vein, and by other operative expedients. None of these laborious efforts worked. We never succeeded in keeping the fistula freely flowing and open, except for a short time. The clinical experience with the young colored woman accomplished more than our best experimental efforts could attain.

The clinical observations cited turn our attention to another problem. Is the protein depletion resulting from loss of thoracic duct lymph simply a loss of blood protein, incident upon free leakage from the blood into intestinal and liver lymph vessels, or has there also been a loss of the protein newly formed in the liver to compensate for depletion? If this last is the case, there is no chance that, in the presence of a freely draining thoracic duct fistula, the patient can gain upon the protein deficit. This statement depends upon acceptance of the liver as the principal, if not the only, depot for the formation of circulating proteins, and evidence available today indicates strongly that this is the fact.

In a recent experiment, CoTui[13] and his associates ligated the thoracic duct in dogs, then bled them severely and followed the concentration of proteins in the blood. These dogs were compared with others, bled similarly, in which the thoracic duct drained normally into the blood. The results were quite decisive. When the thoracic duct was intact, the prehemorrhagic level of blood protein was regained within

two days, at the most. On the other hand, in the dogs with duct ligation, eight days were required to become normal. This time is taken to express the period necessary for opening of free drainage of thoracic duct lymph into other veins, or to effective communication with the right lymphatic duct. The lymph route to the blood, for protein stored or formed in the liver, is, thus, in the same category of essential lymphatic function as is the case for the lymphatics in other parts of the body. This characterization of the accomplishments of the lymphatic system is in accord with what is so readily evident in the frog.

SUMMARY

In the past twenty years, evidence has accumulated showing that the filtrate from the blood capillaries in practically all parts of the body (renal glomerular, choroid, and ciliary vessels being exceptions) contains all of the proteins of the blood in fairly high dilution. This filtrate into the tissues is subject to concentration and to dilution, depending upon capillary blood pressure and other varying contingencies The extravascular protein is, in part, used for tissue cell nutrition, and, in part, enters lymph capillaries, to be moved slowly and rather casually back to the blood. The amount of free fluid in the tissues containing dissolved proteins and other simpler compounds is, normally, very slight, as judged by the amounts of lymph one can collect from regions in which there is no elevation of pressure in the venous capillaries to hinder water absorption. Lymph carrying fat from the intestine, and highly proteinized lymph from the liver, make up the major part of thoracic duct lymph. Evidence is available showing that, in a patient with a very complete thoracic duct fistula, serious loss of blood protein occurred, while, in dogs, bled severely after ligation of the thoracic duct, restoration of blood protein was much delayed. These findings apparently point to the lymph route from the liver as being the means of entrance of stored and newly formed protein into the blood, and place the function of the liver lymphatics in line with that of these vessels in other parts of the body. Only in the villi and lacteals of the small bowel do we find lymphatics specialized for a single novel function: the absorption and delivery of fat toward the blood. Yet, with the emulsified fat in the lacteal lymph, there is also protein, which is derived from capillaries in the intestinal wall, and is probably of great importance for contriving the emulsion which moves slowly into and up the thoracic duct.

The ready entrance of serum albumin, globulin, fibrinogen, and pro-

thrombin into lymphatic capillaries carries with it the certainty of equally easy entrance of other proteins. But the lymph capillaries admit red cells, particles of carbon, bacteria, quartz, and almost anything that can be forced into the tissues. This free entrance of all sorts of particles and infectious agents into lymphatics is, to some degree, minimized by the fact that all lymph, before final return to the blood, passes through a lymph node, an extraordinarily effective mechanical and biological filter, which is also, apparently, an important site of antibody formation. That so elaborate an arrangement of vessels, filters, and producers of protective substances exists, particularly in man, carries with it the implication of danger that such a comparatively isolated system for dealing with concentrations of foreign and often infectious material may fail to provide protection, and become an extensive site of disease.

REFERENCES

1. **Starling, E. H.**
   1896. On the absorption of fluids from the connective tissue spaces. J. Physiol. **19**: 312.
2. **Churchill, E. D., F. Nakazawa, & C. K. Drinker**
   1927. The circulation of body fluids in the frog. J. Physiol. **63**: 304.
3. **Conklin, R. E.**
   1930. The formation and circulation of lymph in the frog. I. The rate of lymph production. Am. J. Physiol. **95**: 79. II. Blood volume and pressure. Am. J. Physiol. **95**: 91. III. The permeability of the capillaries to protein. Am. J. Physiol. **95**: 98.
4. **Drinker, C. K., & M. E. Field**
   1931. The protein content of mammalian lymph and the relation of lymph to tissue fluid. Am. J. Physiol. **97**: 32.
5. **Maurer, F. W.**
   1938. Isolation and analysis of extracellular muscle fluid from the frog. Am. J. Physiol. **124**: 546.
6. **Drinker, C. K., M. E. Field, J. W. Heim, & O. C. Leigh, Jr.**
   1934. The composition of edema fluid and lymph in edema and elephantiasis resulting from lymphatic obstruction. Am. J. Physiol. **109**: 572.
7. **Maurer, F. W., M. F. Warren, & C. K. Drinker**
   1940. The composition of mammalian pericardial and peritoneal fluids. Am. J. Physiol. **129**: 635.
8. **Field, M. E., M. F. Shaffer, J. F. Enders, & C. K. Drinker**
   1937. The distribution in the blood and lymph of pneumococcus type III injected intravenously in rabbits, and the effect of treatment with specific antiserum on the infection of the lymph. J. Exp. Med. **65**: 469.
9. **Whipple, G. H., & S. C. Madden***
   1944. Hemoglobin, plasma protein and cell protein—their interchange and construction in emergencies. Medicine **23**: 215.

* I utilize a single recent reference to the work of Whipple and his group. This is a poor expression of my esteem for the patient and beautiful development of one of the most important advances in physiology made in the past fifteen years. The student desiring to orient himself fully in this most significant group of investigations can do so by utilizing the references appended to the single article I have selected for reference.

10. **Drinker, C. K., M. E. Field, & J. Homans**
1934. The experimental production of edema and elephantiasis as a result of lymphatic obstruction. Am. J. Physiol. **108**: 509.

11. **Crandall, L. A., Jr., S. B. Barker, & D. G. Graham**
1943. A study of the lymph flow from a patient with thoracic duct fistula. Gastroenterology **1**: 1040.

12. **Freeman, L. W.**
1942. Lymphatic pathways from the intestine in the dog. Anat. Rec. **82**: 543.

13. **CoTui, F., I. S. Barcham, & B. G. P. Shafiroff**
1944. Ligation of the thoracic duct and the posthemorrhage plasma protein level. Surg. Gynec. Obst. **79**: 37.

DISCUSSION OF THE PAPER

Dr. Webb:

The aversion to giving serious consideration to the active participation of lymphatic vessels in the propulsion of lymph in the mammal has been continued in the paper read by Dr. Drinker. It is true that, in some of the common laboratory animals, dog, rabbit, and cat, the intrinsic activity of these vessels has not been observed. In his all-inclusive statement concerning mammals, he has chosen to ignore the observations of Heller, Lieben, Florey, Carelton and Florey, Carrier, Webb, Pullinger, and Florey, and Webb and Nicoll on the rhythmical contractions of the lymphatics in the rat, guinea pig and bat. This activity is not only found in the lacteals, but can be observed in the peripheral vessels, as well. If this statement had been confined to the dog, the mammal on which most of his experiments have been conducted, no objection could be raised.

The question of quantitative significance of the inherent propulsive force of the lymphatics has not been adequately analyzed. In those animals whose lymphatics pulsate actively, this propulsive force is the major factor in lymph movement. In its absence, the movement of lymph usually ceases. It is illogical to dismiss this action, as being quantitatively unimportant in the absence of positive evidence that the lymphatics of a few mammals do not show this behavior.

Dr. Melvin Knisely (*Department of Anatomy, University of Chicago, Chicago, Illinois*):

Drs. Landis and Drinker have both spoken about the leakage of protein molecules through the walls of normal capillaries. Dr. Landis summarised evidences from many sources to mean "that the original normal capillary filtrate contains from 0.2 to 0.5 per cent protein." This raises the perplexing problem of how a normal capillary wall can retain most of the plasma protein molecules, yet permit some albumen with a molecular weight of 69,000, some globulin with a molecular weight of 160,000, and even minute amounts of fibrinogen with a molecular weight of 500,000 to pass outward through it, into the tissue spaces, thence to lymphatics. If the capillary wall leaks any of the larger protein molecules, why does it not leak all the smaller protein molecules? How does it retain any protein?

Dr. Landis suggests that the normal capillary wall retains most of the protein molecules of each category, but fails to retain some of each, because: (a) "the molecules of the three important proteins of the blood all have about the same equatorial diameter, differing greatly, however, in length" (all are long and narrow); and (b) "as a working concept, one may imagine the capillary endothelium to consist of a meshwork of pores of many sizes of which a few must have diameters of 38 Ångstrom units or possibly more." According to this concept, as the blood plasma moves lengthwise through a capillary, most of the long, narrow, tumbling, turning protein molecules would strike the inner surface of the endothelium flatwise or at oblique angles. A few would, by chance, be forced end-

wise into the larger pores in the capillary wall and could be forced through, out into the adjacent tissue space. Thus, most protein molecules of each category would be retained and a few of each category escape.

Reduced to its fundamentals, this hypothesis is based upon three concepts: (a) long, narrow protein molecules are (b) presented at all angles to (c) a few pores, each just large enough to admit a narrow protein molecule lengthwise. To me, this hypothesis seems altogether reasonable and plausible; the following material is not presented as an alternative hypothesis, but as an additional hypothesis. It is quite possible that there are two or more sets of mechanism, always in operation, which cause the vascular system to retain most protein molecules, but which still always leak some, so that they are always present in lymph.

Since Professor Krogh's classical studies,[1] it is common knowledge that striated muscles have an intermittent capillary circulation. In 1936, Professor Krogh asked me to study the vascular reactions and changes in circulation in transilluminated muscles, during different phases of muscle activity, such as during rest, rhythmical contractions, tetanic contractions, etc. The following may be taken as a preliminary report of a part of that study:

When a frog's striated muscle goes into a resting phase, most of its arterioles shut off for long periods, and the endothelium of those capillaries through which blood is not flowing becomes progressively anoxic. After a time, probably because of lack of oxygen,[2] they become permeable to blood plasma colloids. Being empty, they do not, of course, leak plasma during the period when no blood is entering them; they cannot lose fluid which they do not contain. However, when the muscle makes one long, strong contraction, or when it goes into rhythmical contractions, the arterioles and capillaries dilate, as a response to the muscle metabolites, and the *first* blood to enter the muscle flows into capillaries which are at that moment permeable to such large molecules that all, or nearly all, of the entering plasma goes out through the capillary walls. This is unmistakable under the microscope, because, as the column of blood passes from arterioles to capillaries, the red cells, moving in a single file or double row, suddenly come progressively closer together, so that, by the time they have traversed from 1/3 to 2/3 of the capillary's length, the capillary contains a moving column of closely packed red cells. There is no space between the red cells nor between the red cells and the endothelial wall for more than very thin films of plasma. The best place to study this is in the ventral surface of the mylohyoid muscles of frogs which are so lightly anesthetized (with urethane) that they respire occasionally, thus using the mylohyoids voluntarily.

The blood pressure available in muscle capillaries certainly is not great enough to force all the water and crystalloids out through endothelium which is impermeable to colloids, and force the retained colloids into almost zero volume of space, against the inwardly directed osmotic "attraction" force of those colloids. This, and the fact that almost the whole plasma volume goes out through the wall, is inescapable proof that most of the plasma proteins go out along with the water and crystalloids.[1] This is proof that, during this phase of the muscle's activity, the capillary endothelium is permeable to almost all of the osmotically active molecules of dissolved blood proteins.

A short time after arterial blood has begun to flow into previously anoxic muscle capillaries, the endothelial walls regain their ability to retain plasma.[2] From then on, if the muscle rests, the red cells do not pack together as they pass along the open capillaries. During rhythmical contractions of the muscle, many of its capillaries do, however, often remain partly permeable to some of the proteins, probably those of lesser molecular weight, for the red cells do come noticeably closer together, but not tight together, as they pass along the capillaries.

These direct observations, made in frogs, of the loss of fluid from anoxic muscle

[1] Krogh, A. The Anatomy and Physiology of Capillaries. Yale University Press. New Haven, Connecticut. Revised Edition. 1929.

[2] Landis, E. M. Micro-injection studies of capillary permeability. 3. The effect of lack of oxygen on the permeability of the capillary wall to fluid and to the plasma proteins. Am. J. Physiol. 83(2): 528–542. 1928.

capillaries, as the muscle begins to contract and as it contracts rhythmically, agree with: (1) Barcroft and Kato's finding that there is an increased loss of fluid from the vessels of dog muscles, during periods of rhythmical contractions;[3] and (2) with White, Field, and Drinker's finding that there is an increase in the rate of lymph return from the legs of unanesthetized dogs, when the dogs begin to exercise the muscles of their legs as they go from rest to walking or running.[4] However, frog blood contains lower concentrations of proteins than mammalian blood, and it is commonly believed that, under a number of conditions, frog capillaries retain proteins less well than mammalian capillaries. Consequently, during muscular exercise, the capillaries of mammalian muscle probably do not normally lose (leak) as large a proportion of the blood plasma flowing through them as do frog muscle capillaries.

For the present purpose, the following points are significant:

(a) There are normal phases of normal physiology during which normal ordinary capillaries are empty, receive no blood, and may be presumed to have partially, to severely, anoxic endothelium.

(b) In striated muscles of the frog, when blood flows through anoxic endothelium, plasma leaks out through the endothelial wall, rapidly enough to be readily detectable by simple observation.

(c) Oxygenated endothelium does not leak blood plasma rapidly enough to be detectable by simple observation.

(d) Endothelium which has not been too anoxic for too long easily regains its ability to retain blood plasma when blood flows through it for a time.

Putting these bits together into a hypothesis, it seems reasonable to assume the following:

(a) In many parts of the body, various numbers of capillaries may have blood flowing through them fast enough to maintain an adequate rate of supply of oxygen molecules to each point along the endothelium of each open capillary, and those capillaries with adequate flow are leaking but little protein.

(b) In many parts of the body, during various phases of normal physiology, some capillaries are shut off for various periods, during which their walls become anoxic.

(c) Whenever one anoxic capillary begins to receive blood again, there is a brief interval, during which its wall is still anoxic and cannot retain proteins, and after which it retains protein molecules increasingly well.

Thus, during normal physiology, the amount of protein passing from blood to tissue spaces, thence to lymph, at any moment, by this mechanism, would depend upon three factors: (a) the numbers of anoxic capillaries just opening up; (b) the summations of the rates of loss through the walls of those just opening; and (c) the duration of the anoxic leaking phases. Thus, this mechanism could easily account for various rates of passage of proteins from blood to lymph.

In summation, two hypotheses have been presented, to account for the fact that, under normal conditions, lymph always contains some protein of large molecular weight, even though the vascular system retains most of the plasma proteins of small molecular weight.

The hypotheses are not antagonistic. The phenomena assumed in each may well be operating simultaneously.

Dr. Nicoll: May I ask whether Dr. McMaster agrees with the view, held by many authors, that there is no lymph flow in limbs at rest?

Dr. McMaster: I have not made direct collections of lymph from human skin, but much of the material included in my paper was presented to bring out the point that there is some movement of lymph, at least in cutaneous lymphatics,

[3] **Barcroft, J., & T. Kato.** Effects of functional activity in striated muscle and the submaxillary gland. Phil. Trans. Roy. Soc. London. B **207**: 149–182. 1915.

[4] **White, J. C., M. Field, & C. K. Drinker.** On the protein content and normal flow of lymph from the foot of the dog. Am. J. Physiol. **103**(1): 34–44. 1933.

of resting limbs of both animals and man. In human skin, dye streamers develop in resting limbs and become longer and more intense under conditions known to increase lymph flow. During the intense reactive hyperemia that follows release of circulatory obstruction, also in the edema of nephrosis, lymph movement seems to be very great, even though the limbs are not moved.

Clearly, lymph flow takes place in the skin of motionless ears of large rabbits, since, in nearly all of many instances, I have obtained a flow of lymph into cannulae placed in lymphatics at the bases of the resting ears. The amount of lymph collected in these experiments was about 5 times as great as that previously obtained by Henry.

As mentioned in the paper, dye introduced intradermally, at the tip of a motionless ear of a rabbit, enters the lymphatics there and moves slowly in the channels toward the ear's base, forming a dye "streamer." In a number of experiments, cannulae were placed in lymphatics at the bases of rabbits' ears, and, after the rate of lymph flow had been determined, dye was introduced into lymphatics at the tips of the ears. The dye streamers which formed progressed to the bases at about the same rate as in the previously mentioned experiments with uncannulated ears. During the progress of the streamers, the rate of lymph flow into the cannulae increased but slightly, or not at all. Clearly, the progress of dye streamers in both cannulated and uncannulated ears yielded a rough approximation of the rate of lymph movement taking place. Further, conditions known to increase lymph flow increased the rate of progress of the streamers and increased the volume of lymph collected from the cannulated ears.

Phenomena have also been reported, in earlier work, indicating a flow of lymph in the superficial channels of the motionless ears of mice. In lymphatic capillaries containing blue dye, one could often see the blue fluid swept aside by a stream of clear lymph entering from regions unaffected by the dye injection.

Since there seems to be some lymph movement in superficial cutaneous channels of motionless animals, I might add a word concerning the treatment of badly damaged limbs by immobilization in plaster casts. It is generally believed that immobilization stops lymph movement, thereby inhibiting the spread of infection by this route. Undoubtedly, lymph flow is inhibited to a great extent by lack of movement, but it is my belief that the immobilization treatment attains its end, not only by this means, but also, in some measure, by the external pressure exerted upon the skin by the plaster casts and paddings. Complete lymph stasis in the skin can be obtained with very slight pressure.

Dr. Clark has reported that there is little or no lymph flow in glass chambers in the ears of rabbits. However, one may suggest that the insertion of the chamber in the ear interferes with the normal progress of lymph movement from the ear's tip to its base, just as I have shown to be the case in the mouse's ear when a burn or incision is present.

capillaries of lymph nodes are capable of large scale absorption of protein, they are not only unique members of the blood capillary system, but at the same time display a miraculous ability to absorb protein so equitably as to make efferent and afferent lymph precisely equal in protein content."[13]

TABLE 1

THE AVERAGE NUMBER OF LEUKOCYTES* FOUND IN PERIPHERAL, INTERMEDIATE, AND CENTRAL LYMPH OF DOGS, CATS, AND RABBITS

Authors	Animals	Leukocytes per cu. mm. of lymph	Monocytes per cent
Peripheral Lymph			
Yoffey and Drinker (1939)	Dogs	550	18
Baker (1932/3)	Cats	699	
Yoffey and Drinker (1939)	Cats	430	11
Nii (1932)	Rabbits	2230	4
Jwaki (1934)	Rabbits	1925	3.4
Okaue and Hojo (1935/6)	Rabbits	2050	
Ehrich and Harris (1942)	Rabbits	3260**	
Ehrich and Harris (1942)	Rabbits	6820***	
Intermediate Lymph			
Goodall and Paton (1905/6)	Dogs	5600	
Baker (1932/3)	Cats	6798	
Nii (1932)	Rabbits	9880	1
Menkin and Freund (1929)	Rabbits	14650	0
Ehrich and Harris (1942)	Rabbits	17000	0-1%
Ehrich and Harris (1942)	Rabbits	54100**	
Ehrich and Harris (1942)	Rabbits	67500***	
Central Lymph			
Rous (1908)	Dogs	6950	
Yoffey (1932/3)	Dogs	9040	
Yoffey and Drinker (1939)	Dogs	7800	
Yoffey and Drinker (1939)	Cats	12000	
Sanders, Florey, and Barnes (1940)	Cats	14300	
Davis and Carlson (1909/10)	Rabbits	24110	
Kindwall (1927)	Rabbits	32606	0.03

* Peripheral lymph often contains a good many polymorphonuclear leukocytes and monocytes, while intermediate and central lymph are practically free of these cells.
** 3–9 days after injection of sheep erythrocytes.
*** 3–9 days after injection of typhoid vaccine.

As the cells in the lymph node undergo active mitotic, as well, probably, as amitotic division, and as this is markedly enhanced in experimental lymphocytosis, it cannot be doubted that lymphocytes are ac-

tually formed in the node. However, this does not necessarily mean that all the lymphocytes of the efferent lymph are produced there. In fact, we have evidence to show that many lymphocytes after completing circulation, return to the various lymphatic tissues and, possibly, thereafter reenter the circulating lymph (cf. p. 827).

If we consider the various figures obtained by Rous,[61] Davis and Carlson,[10] Yoffey,[69, 70] and others, it may be accepted that, in dogs weighing 10 kilograms, the lymph of the thoracic duct may contain about 10,000 lymphocytes per cu. mm. Considering the rate of lymph flow in the duct, it can be calculated, from this figure, that the average output of lymphocytes through this duct amounts to 200 millions per hour or 5 billions per day. Since dogs have about 3,000 lymphocytes per cu. mm. in the circulating blood, and there is about 1 liter of blood in a dog weighing 10 kilograms, it can be estimated that the circulating blood of this animal contains not more than 2½ billion lymphocytes or half the number which every day enters the blood through the thoracic duct. This means that the lymphocytes stay in the circulating blood not longer than half a day; in other words, the blood lymphocytes are replaced at least twice a day.

In cats and rabbits, Sanders, Florey, and Barnes[62] have found even higher figures. In cats, the lymphocytes of the blood are replaced at least 2 to 3 times a day, while in rabbits, they are replaced at least 5 times a day.

The validity of these calculations is borne out by other observations. After ligation of the thoracic duct and the trunks of both cervical lymphatics, in rabbits, Bunting and Huston[6] observed an immediate drop in the number of circulating lymphocytes, from 3000 to 900 per cu. mm., in one animal, and from 5000 to 300 per cu. mm., in another. Similarly, Minot and Isaacs,[49] after transfusing a patient with 450 cc. of blood containing 85,000 lymphocytes per cu. mm., first observed an immediate rise in the number of circulating lymphocytes, from 900 to 3,740 per cu. mm. This was followed by a subsequent fall to 960 per cu. mm., within 24 hours.

After having shown that the lymphocytes rapidly disappear from the circulating blood, let us investigate where these cells go, and what their fate may be.

THE FATE OF THE LYMPHOCYTES

It has been claimed by some investigators that it is the fate of the lymphocytes to be transformed into monocytes, granulocytes, or eryth-

rocytes. There is no time here to review the numerous contributions published on this issue. Most of these papers have been critically reviewed on an earlier occasion.[19] It is necessary, however, to discuss briefly some recent publications of Bloom and Yoffey, for the conclusions presented in these papers have apparently been accepted in some quarters.

Bloom[8] has claimed that he succeeded in growing monocytes as well as granulocytes in tissue cultures from lymphocytes of the thoracic duct. Hall and Furth,[30] and Medawar,[47] who repeated this work, also observed the growth of monocytes. However, when counting the cells at different intervals after explantation, Hall and Furth found that the increase in monocytes was merely a relative phenomenon, caused by rapid disintegration of the lymphocytes; while Medawar observed that the occasional appearance of granulocytes in cultures of thoracic duct lymphocytes was due to improper technique of collecting lymph. Lewis,[45] and Ebert, Sanders, and Florey[14] have never observed transformation of lymphocytes into other cells, though individual cells were watched for many hours.

Yoffey[18] has contended that the relative rates of disappearance of lymphocytes from the blood and of new formation of erythrocytes in the bone marrow conformed to the old idea of Jordan[39] that the fate of the lymphocytes was to enter the bone marrow, where they were transformed into erythrocytes. Recently, Yoffey and Parnell[72] have claimed that 12.5% of all the nucleated cells of the bone marrow were lymphocytes, and that the total number of lymphocytes in the marrow equalled the average daily output of these cells through the thoracic duct. Since 60 to 70% of all nucleated cells of the bone marrow are granulocytes, the ratio of lymphocytes to nucleated red cells, if Yoffey's claim were the fact, would amount to 1:2 or 1:3. Such a ratio, however, is not observed, under usual conditions.

The chief objection to Jordan's idea springs from hematological experience and pathological observations in various blood diseases. For instance, Sjoevall[65] has shown that continued bleeding leads to atrophy of the lymphatic tissue and to lymphocytopenia, while the bone marrow reveals greatly increased erythrocytopoietic activity. On the other hand, Tuta[67] found no increase in the number of lymphocytes in the bone marrow in marked experimental lymphocytosis caused by *Hemophilus pertussis*.

It has recently been shown by us[31] and by Dougherty, Chase, and White[12] that the lymphocyte has something to do with antibody pro-

duction and possibly plays an important role in the γ-globulin household. This discovery should further discredit the theory that the lymphocyte is an immature blood cell capable of transformation into monocytes, granulocytes, erythrocytes, or other blood cells.

According to a second theory to be mentioned here, the lymphocyte ultimately disappears from the circulation by rapid disintegration within the blood stream. If this assumption were fact, the lymphocytes in the rabbit would have to disappear within 4 to 5 hours, because they are replaced at least 5 times a day. It has been shown by Heineke[33] that in lymphatic tissue, after irradiation, the lymphocytes begin to disintegrate as soon as two hours after, and chromatin fragments are abundant after 4 to 8 hours. However, digestion of the fragments is accomplished only after 24 to 36 hours, which is 6 to 7 times the circulation time of the lymphocyte.

A third and more important theory of the fate of the lymphocyte, which should be discussed here, is that which was proposed by Bunting and Huston.[6] When collecting fluid from close to the mucosa, these authors found, in the duodenum of rabbits, 100 lymphocytes per cu. mm. of fluid; in the ileum, 480; and in the appendix, 700. In fact, "if the 40 or 50 lymphocytes which may be counted between the epithelial cells of the tip of a duodenal villus in a 10 mikra section are multiplied by its probable relation to the surface area of the intestine, there is no difficulty in accounting for all the lymphocytes that disappear from the circulation."

These observations of Bunting and Huston have been confirmed by Jassinowsky,[36] who, by an irrigation method, found an emigration rate of 4000 lymphocytes per minute per sq. cm. in the duodenum of rabbits, 7100 in the ileum, and 14500 in the appendix. Similarly, by counting the leukocytes in the saliva, Isaacs and Danielian[37] observed marked increase in the emigration rate of granulocytes and lymphocytes in lymphoblastoma; in fact, in aleukemic leukemia the number of lymphocytes per cu. mm. of saliva was sometimes larger than that in the blood. Ohno,[55] finally counted 345 to 460 mononuclear cells per cu. cm. before meals, and 3840 to 3050 after meals, in the juice from a fistula of the lower ileum of a man.

It should be mentioned here that Erf,[23] after removing spleen and gastrointestinal tract in rabbits, observed a rapid fall in the total number of lymphocytes in the blood, from an average of 1700 per cu. mm. to one of 300 per cu. mm. He further observed that, after intravenous injection of 200 to 300 million lymphocytes into such an animal, the

number of circulating lymphocytes did not rise, as would have been expected if the theory of Bunting and Huston were correct. Though these experiments deserve further consideration, they are hardly qualified to discredit the direct observations of Bunting and Huston, Jassinowsky, Isaacs and Danielian, and Ohno. These latter observations leave little doubt that at least some lymphocytes leave the blood through the gastrointestinal tract.

A fourth theory concerning the fate of the lymphocytes is the one advocated by Sjoevall.[65] According to this theory, the lymphocytes leave the blood through the peripheral capillaries to enter the peripheral lymph vessels, and thus return to the lymph nodes. It is true that the fate of a good many lymphocytes is satisfactorily explained by this mechanism. The total number of lymphocytes thus accounted for, however, is small, for the peripheral lymph usually contains less than *one-tenth of the number of lymphocytes* contained in the intermediate lymph (TABLE 1).

The last and possibly most important theory of the fate of the lymphocyte to be considered here, is that which has been proposed by Heiberg.[32] According to this theory, many lymphocytes return through the blood capillaries directly to the lymphatic tissue, to be destroyed within the so-called germinal centers. If this theory were correct, lymphatic tissue would be not only the birth place, but also the graveyard of lymphocytes. That this is probably true, is borne out, as will be seen, by the various events that may be observed within the lymphatic tissue, under various spontaneous and experimental conditions.

THE ROLE OF THE LYMPHATIC TISSUE IN THE CIRCULATION OF THE LYMPHOCYTE

By lymphatic tissue, we mean organized lymphoid tissue, characterized by the presence of lymphatic nodules consisting chiefly of lymphocytes. Lymphatic tissue, in this sense, we find spread through the entire organism. The bulk of this tissue is located in the lymph nodes, in certain mucous membranes, and in the spleen.

According to their morphology, time of appearance, and functional significance, three types of lymphatic nodules may be distinguished. These have, in the past, been called follicles, primary nodules, secondary nodules, germinal centers, functional centers, reaction centers, pseudosecondary nodules, etc. Since these terms have not been uniformly

applied and some have given rise to confusion, it is suggested that most of them be discarded in favor of the terms, primary, secondary, and tertiary nodules.

Primary nodules (the solid secondary nodules of Groll and Krampf[29]) consist nearly exclusively of small lymphocytes (PLATE 5). They are the first lymphatic nodules to appear, and, in man, have been found as early as in the 22nd fetal week.[16] The diameter of these nodules does not usually exceed 0.3 mm.

Secondary nodules (the germinal centers of Flemming[25]) are well circumscribed, pale areas of medium-sized and large cells which are mostly interpreted as hemocytoblasts, lymphoblasts, or special lymphocytes (PLATE 6). Secondary nodules appear only after birth. Their diameter may be 0.75 mm. or larger.

In most cases, secondary nodules are partially or totally surrounded *by a sharply delimited marginal zone of small lymphocytes.* Since this zone may be absent, it cannot be considered an essential part of a secondary nodule. In fact, the presence of a marginal zone is mostly interpreted to indicate that the secondary nodule arose, primarily, within a primary nodule.

A special feature of secondary nodules is the presence of regularly distributed macrophages containing globular masses of chromatin (the tingible bodies of Flemming (PLATE 6, right). The latter are mostly disintegrating nuclei of lymphocytes, as was shown by Heineke through irradiation experiments, many years ago.[33]

Tertiary nodules (the pseudosecondary nodules of Ehrich[15]) are large, fairly well circumscribed nodes of lymphoid tissue, frequently measuring 3 mm., or more (PLATE 7). Surrounding these, toward their surface, there are usually several small primary or secondary nodules. Tertiary nodules may be observed even in fetal life.[16]

Primary and secondary nodules, on the one hand, and tertiary nodules, on the other, differ greatly in their blood supply. Primary and secondary nodules are usually supplied by an arteriole and the arterial portions of its capillaries, while veins are completely absent (PLATE 8).* Tertiary nodules, on the other hand, contain both arteries and veins, the veins being characterized by a peculiar endothelial lining consisting of densely crowded, rather high endothelial cells resembling cuboidal epithelium (PLATE 9). Indeed, these vessels resemble glands so closely that a noted pathologist recently described them as "glandular inclusions."[66] Since the arteries contain only few lymphocytes, while, par-

* For the blood vessels of the lymphatic nodules, see **Calvert,**[7, 8] **Ono & Mijazaki,**[55] **Dabelow,**[9] **Fischer,**[24] and **Ono.**[57]

Considering the function of the lymphatic nodules, it seems that the tertiary nodules are merely the products of marked lymphocytopoiesis and equipped with special facilities to dispatch lymphocytes into the lymph and, possibly, directly into the blood.

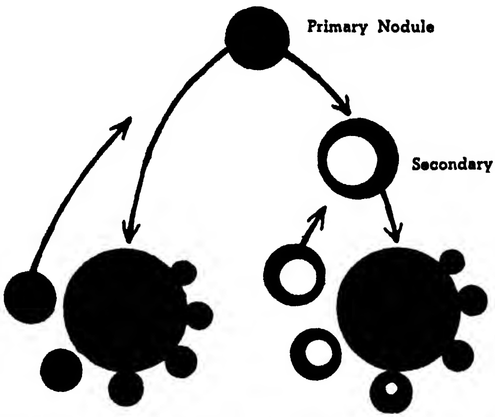


FIGURE 3. Diagram showing the fetal (left) and post-fetal (right) cycles of the lymphatic nodules.

The primary nodules may merely be foci of active lymphocytopoiesis resulting in the production of small lymphocytes, though it should not be overlooked that mitotic figures are conspicuously absent. It is equally possible, however, that primary nodules arise, not through new formation, but through the return to the lymphatic tissue, by way of the capillaries, of lymphocytes that have completed circulation. The latter contention, however, seems to be inconsistent with the well-known fact that leukocytes emigrate from venules, rather than from capillaries.

As to the function of the secondary nodules, three theories deserve special consideration. The oldest of these is that of Flemming,[25] who claimed that the secondary nodules were the birth places or germinal centers of the small lymphocytes. This theory was based chiefly on the presence in the nodules of many mitotic figures. The arguments for and against this theory have been discussed elsewhere.[18] It may suffice to mention here that the large and medium-sized lymphoid cells of the secondary nodules cannot very well be regarded as reticuloendothelial cells, because they do not store vital dyes; nor are they likely to be hemocytoblasts, since, if other blood cells appear in them, they

arise, not from the lymphoid cells, but from elements attached to the capillaries.[19] In fact, all evidence that can be mustered seems to show that they are lymphoblasts. Therefore, lymphocytopoiesis has continued to be regarded at least as one function of the secondary nodules.

The second theory to be mentioned here is that of Heiberg,[32] who believed that the secondary nodules were the graveyards of lymphocytes, rather than their birthplaces. Heiberg based this view on the presence in these centers of large numbers of macrophages, filled with the fragments of dead lymphocytes (PLATE 6, right). If we consider that one secondary nodule may contain several hundred of such macrophages, and each macrophage may contain the fragments of 20 or more lymphocytes, we can see that there may be many thousand disintegrating lymphocytes in one secondary nodule. If we further consider the large number of secondary nodules that are usually present in the numerous lymph nodes, in the gastrointestinal tract and elsewhere, it becomes clear that a large portion of the lymphocytes that are lost from the blood may be accounted for by this mechanism. This estimation is consistent with the findings of Kindred,[42] who, by counting the disintegrating lymphocytes in the various lymphatic tissues, came to the conclusion that, in rats, the number of lymphocytes destroyed by this method amounted to over 1 million per hour per 100 gm. of body weight.

The "graveyard" theory is consistent, also, with the results of experimental lymphocytosis. When, following stimulation of lymphocytopoiesis by typhoid vaccine or sheep erythrocytes, the output of lymphocytes through the efferent lymph vessel was compared with the morphologic changes that took place within the node, it was found that, during the first 3 days when the number of lymphocytes in the efferent lymph rose from an average of 15,000 to one of 60,000, the pre-existing lymphatic nodules became markedly enlarged, lost their identity, and changed into a rather diffuse mass of lymphoid tissue, showing many mitotic figures.[20] Only 5 to 6 days after the antigen was injected, or 2 to 3 days after lymphocytopoiesis had reached its peak, did secondary nodules and fragments of lymphocytes make their appearance. After 6 to 21 days, there were numerous, large, confluent, secondary nodules showing many mitotic figures as well as numerous fragments of disintegrating lymphocytes (TABLE 2). The secondary nodules and fragments thus appeared after a time which has generally been found to be the life span of the lymphocytes.

It should be obvious that the "graveyard" function of the secondary nodules does not exclude a "birthplace" function. In fact, this simul-

TABLE 2

THE TIME RELATIONSHIP BETWEEN LYMPHOCYTOPOIESIS, ANTIBODY FORMATION, ACTIVITY OF SECONDARY NODULES, AND DISINTEGRATION OF LYMPHOCYTES IN THE REGIONAL LYMPH NODE, FOLLOWING SUBCUTANEOUS INJECTION OF TYPHOID VACCINE AND SHEEP ERYTHROCYTES*

Duration of experiment days	Rabbit No.	Efferent lymph		Serum antibody** titer	Lymph node		
		No. of lymphocytes per cu. mm.	Antibody** titer		Weight gm.	Secondary nodules	Fragments of disintegrating lymphocytes
Normal Controls							
0	13	18,000	0	0	.2	some	a good many
..	14	16,000	0	0	.2	some	a good many
Typhoid Vaccine							
1	24	.....	...	....	.35	some	a good many
2	73	.....	...	0	.65	none	none
3	34	44,900	2	8	.65	none	none
	97	172,800	0	4	.85	none	none
4	16	133,500	48	16	.5	none	none
	35	59,600	64	128	.5	none	none
5	17	86.000	64	192	.9	early	some
	36	35,000	192	256	.85	many early	some
	93	36,800	64	128	.85	early	some
6	18	.....	...	1024	.7	large	numerous
	37	58,500	192	768	.55	many	many
	87	81,900	128	256	1.1	many large	numerous
	22	47,000	12	32	1.25	large confluent	numerous
9	26	47,200	64	512	1.05	large	numerous
	59	.....	...	....	.9	many large	a good many
	98	58,300	64	256	1.05	large	numerous
14	60	21,300	64	192	.4	many large	numerous
	99	61,600	32	128	1.45	large	numerous
21	30	29,800	64	256	.6	large confluent	numerous
	100	86,200	32	64	.75	large	quite numerous
29	101	41,650	8	12	.55	large	numerous

* The figures presented here have partly been published. (J. Exp. Med. 76: 340. 1942.) Similar results were obtained with egg albumen and dysentery bacilli.

** The antibodies recorded here are agglutinins versus typhoid vaccine and hemolysins versus sheep erythrocytes.

TABLE 2 (continued)

THE TIME RELATIONSHIP BETWEEN LYMPHOCYTOPOIESIS, ANTIBODY FORMATION, ACTIVITY OF SECONDARY NODULES, AND DISINTEGRATION OF LYMPHOCYTES IN THE REGIONAL LYMPH NODE, FOLLOWING SUBCUTANEOUS INJECTION OF TYPHOID VACCINE AND SHEEP ERYTHROCYTES*

Duration of experiment days	Rabbit No.	Efferent lymph		Serum anti-body** titer	Lymph node		
		No. of lymphocytes per cu. mm.	Anti-body** titer		Weight gm.	Secondary nodules	Fragments of disintegrating lymphocytes
				Sheep Erythrocytes			
1	24	56,800	0	0	.2	some	a good many
2	38	41,200	0	0	.2	none	none
3	34	82,400	0	0	.25	none	none
	39	87,400	0	0	.45	none	none
4	16	44,000	0	0	.2	a few early	a few
	35	84,800	16	4	.25	a few early	a few
	94	22,400	0	4	.3	none	none
5	17	55,000	128	64	.35	none	none
	36	59,600	192	64	.3	many early	some
6	37	65,000	512	256	.3	many	many
	74	30,600	128	256	.3	many large	numerous
	28	51,100	64	512	.4	large	numerous
9	26	100,400	48	1025	.4	large	numerous
	40	32,200	32	256	.3	many	numerous
	27	78,200	64	512	.45	large	numerous
14	60	35,100	32	512	.25	many large	numerous
	41	15,000	16	32	.25	many	numerous
	33	16,500	48	256	.25	many large	numerous
21	30	23,000	4	64	.2	many	quite numerous
	42	27,200	4	8	.3	large	numerous
29	101	24,600	0	16	.3	many	many

* The figures presented here have partly been published (J. Exp. Med. 76: 340. 1942). Similar results were obtained with egg albumen and dysentery bacilli.

** The antibodies recorded here are agglutinins versus typhoid vaccine, and hemolysins versus sheep erythrocytes.

taneous occurrence may be regarded as a good example of the old Italian idea that disintegrating cells liberate specific substances which promote the growth of their own kind (necrohormones).

The third theory which should be discussed here is that of Hellman:[35, 36] namely, that the secondary nodules should be regarded as reaction centers which act as detoxifying organs, or organs of antibody formation. This theory was chiefly based on the observation that secondary nodules are usually produced as the result of inflammation of moderate intensity, while marked inflammation results in their necrosis. It was supported by the fact that the secondary nodules appear only some time after birth, when the infant has lost the protection conferred on him by his mother. It was consistent also with the observation of Hellman's pupil, Glimstedt,[27] that guinea pigs raised in the absence of bacteria fail to develop secondary nodules. Though these observations may justify calling the secondary nodules centers of reaction, they do not necessarily indicate that they are organs of detoxification or antibody formation. It is true that, during immunization, the rise in antibody titer in the serum has been found to parallel the activity of secondary nodules in the spleen (Ehrich and Voigt,[22] Oesterlind[54]). However, a more detailed analysis of this relationship revealed that the secondary nodules reached the peak of their development only after the peak of antibody formation had passed, and they continued to be active long after the antibody titer had begun to decline (Ehrich and Harris[20]). Similarly, the fragmentation of lymphocytes was found to be a sequel, rather than a precursor, of antibody formation, and, in fact, fragmentation of lymphocytes and activity of secondary nodules appeared to be irresolvably tied up with one another (TABLE 2). These observations do not necessarily speak against Hellman's theory, but it cannot be denied that they were satisfactorily explained, if the reaction expressed by the secondary nodules were nothing but one of lymphocytopoiesis and of lymphocyte destruction.

We may say, then, that the evidence available at present seems to show that the role of the lymphatic tissue in the circulation of the lymphocyte is the birthplace and graveyard of this cell. After having been formed in the lymphatic tissue, the lymphocyte passes through the lymphatics and, possibly, also through the veins of this tissue, into the blood stream, where it circulates for several hours. After this period, some leave the blood through the mucous membranes of the gastrointestinal tract; others, probably the majority, return to the lymphatic tissue through the blood vessels contained in this tissue, and

in the lymph nodes, also through the peripheral lymph vessels. It is likely that many of these lymphocytes go on and return to the blood stream, and in fact, this cycle may be repeated several times. When they finally approach death, they are taken up by the macrophages, while passing through a secondary nodule. The evidence available, at present, seems to indicate that lymphocytes complete their life cycle within a few days.

After having shown that we have no reason to assume that the lymphocyte is merely an undifferentiated hemocytoblast, and that all the evidence which we can muster points to the independent life cycle and special function of this cell, let us investigate what this function may be.

THE FUNCTION OF THE LYMPHOCYTE

The function of the lymphocyte has long been a mystery. Indeed, until recently, most of us agreed with the eloquent statement of Rich,[60] made nine years ago, that "literally nothing of importance is known of these cells other than that they move and that they reproduce themselves." In fact, "the complete ignorance of the function of this cell is one of the most humiliating and disgraceful gaps in all medical knowledge. Produced daily in enormous numbers by a mass of distributed lymphoid tissue which, if gathered together, would form one of the most imposing organs of the body, these cells must undoubtedly serve the body in a most essential way, and yet no information is possessed regarding their function, apart from speculation, based on evidence that is equivocal, to say the least."

There was, however, circumstantial evidence that the function of the lymphocyte was concerned with antibodies. This evidence was partially of a pathological nature[5, 11] and partially experimental.[34, 52] That it was not considered seriously, was probably due to the great prestige of the reticuloendothelial theory of antibody formation.

The *pros* and *cons* of the reticuloendothelial theory have been discussed elsewhere.[21] It has been pointed out that little has been added to what was known to Metschnikoff, 60 years ago. It has merely been shown that the macrophage, like the granulocyte, engulfs and digests formed antigenic material, and it has been revealed that proper blockade of the reticuloendothelial system may interfere with the process of antibody formation. The products of digestion of the macrophage have never been identified. There is no evidence to show that the antibodies are mere products of this digestion.*

*Since completion of this presentation, it has been shown that the macro-

That antibodies may be formed in lymph nodes, was first shown by McMaster and Hudack.[46] If two different antigens were injected, one into each ear of mice, the corresponding antibody appeared first in the lymph node of the same side. As to the cells that produced the antibodies, McMaster and Hudack expressed no opinion; the results of their experiments were compatible with both the reticuloendothelial and lymphocytic theory of antibody formation.

The experiments of McMaster and Hudack were followed by our experiments,[20] in which the production of antibodies in a lymph node was compared with the cellular changes that took place in this node,

TABLE 3

ANTIBODY AND LYMPHOCYTE FORMATION IN THE POPLITEAL LYMPH NODE 6 DAYS AFTER SUBCUTANEOUS INJECTION OF TYPHOID VACCINE INTO THE LEFT FOOT, AND SHEEP ERYTHROCYTES INTO THE RIGHT*

	Antibody titer					No. of lymphocytes per cu. mm. of lymph	
	Foot tissue	Afferent lymph	Lymph node	Efferent lymph	Blood serum	Afferent lymph	Efferent lymph
Left foot (Agglutinin titer *vs. E. typhosa*)	16	16	1024	64	512	2,700	80,600
Right foot (Agglutinin titer *vs.* erythrocyte)	..	0	192	32	64	3,200	72,100
(Hemolysin titer *vs.* erythrocyte)	..	12	1024	768	512	3,200	72,100

* The figures presented in this table have been taken from Table I, J. Exp. Med. 76: 338. 1942.

as well as with the output of antibodies and cells through the efferent lymph vessel of this node. Using the hind foot of the rabbit, first introduced into lymphology in 1929,[17] we compared antibody titers and cellular response in the pad of the foot (which was the site of injection of antigen); the lymph contained in the afferent lymph vessel; the popliteal lymph node (the only node regional to the site of injection); the efferent lymph; and the serum (TABLE 3). Antibodies first appeared in the efferent lymph 2 to 4 days after the injection of the antigen, and reached their highest titer after 6 days (TABLE 2). In all experiments,

phage does not synthesize antibody (Ehrich, W. E., T. J. Exp. Med. 83: 373. 1946), but, like the granulocy solving particulate antigen (Harris, T. N., & W. E. press).

it was found that the antibody titer was higher in the efferent lymph; in some cases, the concentration was 100 times greater than that found in the afferent lymph. The production of antibody in the lymph node was preceded and accompanied by a rise in the output of lymphocytes in the efferent lymph, from an average of 17,000 per cu. mm. to 60,000 per cu. mm. or more (TABLE 2). The rise in the output of lymphocytes was preceded and accompanied by greatly increased lymphocytopoiesis everywhere in the lymph node. Numerous lymphoblasts made their appearance, there was marked mitotic activity, and the node became rapidly filled with a diffuse lymphoid tissue. The cellular response during antibody formation was chiefly lymphocytic, and, in most experiments, reticuloendothelial cells were conspicuously absent (PLATE 13). These observations seemed to show that the lymphocyte was a factor in the formation of antibodies.

TABLE 4

THE COMPARISON OF ANTIBODY TITER IN THE LYMPH PLASMA AND LYMPHOCYTE FRACTIONS OF EFFERENT LYMPH DURING ANTIBODY FORMATION*

Antigen	Time after injection of antigen	Rabbit No.	Lymph	Cells per cu. mm. of undiluted lymph	Anticoagulant	Cell volume	Titer of lymph plasma	Titer of lymph cell extract	Ratio of titers
	Days		cc.		cc.	cc.			
Sheep erythrocytes	5	15 Rt.†	1.00	43,700	0.15	0.0080	64	512	1:8
		15 Lt.	0.30	65,350	0.10	0.0046	32	512	1:16
	7	5 Rt.	0.55	46,100	0.10	0 0049	64	128	1:2
		5 Lt.	0.40	47,500	0.10	0.0038	64	96	1:1.5
Typhoid bacilli	5	16 Rt.	1.00	60,950	0.15	0.0118	1024	4096	1:4
		16 Lt.	1.00	57,500	0.15	0.0114	1024	6144	1:8
	7	19 Rt.	0.45	67,350	0:15	0.0122	2048	4096	1:2
		19 Lt.	0.90	51,350	0.15	0.0174	1024	4096	1:4

* The figures presented in this table have been taken from Table I, J. Exp. Med. 81: 75. 1945.

† Rt. refers to the right leg, Lt. to the left leg.

After these preliminary experiments, Harris, Grimm, Mertens, and Ehrich[31] proceeded to the study of a more isolated system. We collected efferent lymph from antibody-producing popliteal lymph nodes, [illegible] compared the concen-

tration of antibodies within the two fractions. It was found (TABLE 4) that the lymphocyte fraction in many instances contained from 8 to 16 times as much antibody as the lymph plasma fraction.*

While our paper was in press, Dougherty, Chase, and White[12] published a brief report showing that, after repeated intra-abdominal injections with sheep erythrocytes, the lymphoid cells from minced pools of the inguinal, axillary, cervical, and mesenteric lymph nodes and thymuses of groups of several mice contained about twice as much antibody as the serum of these animals. Recently, White and Dougherty,[68] and Kass[41] have shown that lymphocytes from minced pools of lymph nodes also contain normal γ-globulin.

These various observations offered several possible interpretations. It was conceivable that the lymphocytes either absorbed or adsorbed the globulins, or, during reproduction, incorporated them into their cytoplasm. In order to test this theory, we have taken normal lymphocytes and incubated them with antibody-containing lymph plasma, and we have injected antibody-containing serum into living lymph nodes and, several hours later, have removed the lymph and determined antibodies in lymphocytes and lymph plasma. In no case have we been able to show absorption, adsorption, or incorporation of antibodies by the lymphocytes.

Another possibility that presented itself was that the lymphocytes were the primary sources of our globulins. It was noted that the ratio between lymphocyte titer and lymph plasma titer was greatest on the 5th day after injection of the antigen, when antibody formation in the lymph node was greatest, while on the 7th day the ratio had dropped considerably (TABLE 4).[31] Moreover, it was observed that the cytoplasm of the rapidly dividing lymphocytes in the lymph node, during antibody formation, was deeply basophilic (PLATE 13), indicating the presence of large quantities of ribose nucleic acid, and the synthesis of protein in these cells (Mirsky[50]). On the other hand, the reticuloendothelial cells showed only a few basophilic stipplets or no basophilia at all. Both these observations were consistent with a primary appearance of antibodies in the lymphocytes.

It is true that the lymphocytes, unlike the granulocytes or the macrophages, are not phagocytic and, therefore, cannot absorb bacteria or other corpuscular antigens. But who can claim that phagocytosis and antibody formation are the same process? It is possible that the granulocytes and macrophages function merely through dissolving

* This work was presented first before the Pathological Society of Philadelphia, October 18, 1944.

corpuscular antigen, and that products of this "digestion" induce the lymphocytes to produce antibodies. If this view were correct, bacteria and other formed antigens would become effective in the "training" of lymphocytes to produce antibodies instead of normal γ-globulin, only after having been prepared by phagocytes. It is not conceivable that microscopically visible organisms could effect "training" of lymphocytes by direct action.

The lymphocytic theory of antibody formation is consistent with the modern views of Alexander[1] and Sevag.[63] The effective forces could create in the lymphocyte a new reaction through modification of a gene, *i.e.*, through mutation. This would go well with the observation that antibody formation in the lymph node is tied up with mitotic division of lymphocytes, since it has long been noted that, during mitosis, cells are particularly susceptible to modifying influences. Alternatively, these forces could act in the lymphocyte as catalysts and, by accelerating thermodynamically possible reactions, directly induce antibody formation. Both these views have been found to be in accord with the chemical concepts of Breinl and Haurowitz,[4] Mudd,[51] and Pauling[59]: that antibody molecules are distinguished from normal globulin by a different order of the amino acid residues in the polypeptide chains, or by a different configuration of the chain, *i e.*, the way that the chain is coiled in the molecule (Sevag[63]).

It is hardly necessary to mention that the lymphocytic theory of antibody formation is consistent with the modern view that antibodies are formed *de novo.*[26] It would well explain the fact that a relatively small quantity of antigen can give rise to many times its units of antibody, as it explains the observation that, when an antigenic substance is labelled in such a fashion that it can be readily identified, no trace of it is found in the corresponding antibody.[26] The lymphocytic theory would also explain why blockade of the reticuloendothelial system can interfere with antibody formation; it could disturb the preliminary "digestion" of formed antigen and, thus, interfere with the appearance of the forces that induce the production of antibody, instead of normal γ-globulin.

Though our various observations were well explained by the lymphocytic theory of antibody formation, there was still another possible interpretation, namely, that the globulins were primarily formed by the plasma cells contained in the lymph nodes, and either they were "labeled" with the aid of the lymphocytes, or, at the time of antibody determination, they existed in the form of shed cytoplasm and,

therefore, were spun down with the lymphocytes each time the preparations were centrifuged. These possibilities deserve serious consideration, as hyperglobulinemia in man is usually associated with an increase in the number of plasma cells in bone marrow and elsewhere, while, in lymphatic leukemia and related conditions, the globulins are mostly within normal levels. The role of the plasma cells is under investigation at present. The close coexistence of lymphocytes and plasma cells in inflammation suggests that both are engaged in antibody production, either in different phases of the same process, or in the production of different antibodies.

Whatever conclusion will finally prevail, we cannot escape the impression that the lymphocyte is not merely an undifferentiated hemocytoblast, but that it plays a definite role in the circulation of the lymph, and in the protein metabolism which is so closely connected with this circulation.

BIBLIOGRAPHY

1. **Baker, R. D.**
   1932–1933. Anat. Rec. **55**: 207.
2. **Biedl, A., & A. V. Decastello**
   1901. Arch. Physiol. **86**: 259.
3. **Bloom, W.**
   1937. Anat. Rec. **69**: 99.
4. **Breinl, F., & F. Haurowitz**
   1930. Zeitschr. f. phys. Chem. **192**: 45.
5. **Bunting, C. H.**
   1925. Wisc. Med. J. **24**: 305.
6. **Bunting, C. H., & J. Huston**
   1921. J. Exp. Med. **33**: 593.
7. **Calvert, W. J.**
   1937. Anat. Ans. **13**: 174.
8. **Calvert, W. J.**
   1901. Bull. Johns Hopk. Hosp. **12**: 177.
9. **Dabelow**
   1936. Verh. Anat. Ges. **43**: 187.
10. **Davis, B. F., & A. J. Carlson**
    1909–1910. Am. J. Physiol. **25**: 173.
11. **Doan, C. A., & F. R. Sabin**
    1930. J. Exp. Med. **52**: 113.
12. **Dougherty, T. F., J. H. Chase, & A. White**
    1944. Proc. Soc. Exp. Biol. & Med. **57**: 295.
13. **Drinker, S. K., & J. M. Yoffey**
    1941. Lymphatics, Lymph and Lymphoid Tissue. Harvard University Press.
14. **Ebert, R. H., A. G. Sanders, & H. W. Florey**
    1940. Brit. J. Exp. Path. **21**: 212.

15. **Ehrich, W.**
    1929. Am. J. Anat. **43**: 347.
16. **Ehrich, W.**
    1929. Am. J. Anat. **43**: 385.
17. **Ehrich, W.**
    1929. J. Exp. Med. **49**: 347.
18. **Ehrich, W.**
    1931. Beitr. Path. Anat. **86**: 267.
19. **Ehrich, W.**
    1934. Ergebn. allg. Path. **29**: 1.
20. **Ehrich, W., & T. N. Harris**
    1942. J. Exp. Med. **76**: 335.
21. **Ehrich, W., & T. N. Harris**
    1945. Science **101**: 28.
22. **Ehrich, W., & W. Voigt**
    1934. Beitr. Path. Anat. **93**: 348.
23. **Erf, L. A.**
    1940. Am. J. Med. Sci. **200**: 1.
24. **Fischer, H.**
    1937. Ztschr. Mikr. Anat. Forschg. **41**: 229.
25. **Flemming, W.**
    1885. Arch. Mikr. Anat. **24**: 50.
26. **Gay, F. P.**
    1935. Agents of Disease and Host Resistance. Thomas Springfield, Ill.
27. **Glimstedt, G.**
    1936. Acta Path. et Micro. Scand. Suppl. **30**.
28. **Goodall, A., & D. N. Paton**
    1905–1906. J. Physiol. **33**: 20.
29. **Groll, H., & F. Krampf**
    1920–1921. Zentralbl. Path. **31**: 145.
30. **Hall, J. W., & J. Furth**
    1938. Arch. Path. **25**: 46.
31. **Harris, T. N., E. Grimm, E. Mertens, & W. Ehrich**
    1945. J. Exp. Med. **81**: 73.
32. **Heiberg, K. A.**
    1922–1923. Arch. Path. Anat. **240**: 301.
33. **Heineke, H.**
    1905. Mitt. Grenzgeb. Med. & Chir. **14**: 21.
34. **Hektoen, L.**
    1915. J. Inf. Dis. **17**: 415.
35. **Hellman, T. J.**
    1918–1919. Upsala Laek. Foerh. **24**: 283.
36. **Hellman, T. J.**
    1921. Beitr. Path. Anat. **68**: 333.
37. **Isaacs, R., & A. C. Danielian**
    1927. Am. J. Med. Sci. **174**: 70.
38. **Jassinowsky, M. A.**
    1925. Frankf. Ztschr. Path. **32**: 238.
39. **Jordan, H. E., & C. C. Speidel**
    1923. Anat. Rec. **26**: 223.

40. **Jwaki, Y.**
1934. Arb. 3. Abt. Anat. Inst. k. Univ. Kyoto (Series D) **4**: 7.
41. **Kass, E. H.**
1945. Science **101**: 337.
42. **Kindred, J. E.**
1942. Am. J. Anat. **71**: 207.
43. **Kindwall, J. A.**
1927. Bull. Johns Hopkins Hosp. **40**: 39.
44. **Krumbhaar, E. B.**
1942. Lymphatic Tissue, Problems of Ageing (E. B. Cowdry). Williams and Wilkins. Baltimore, Md.
45. **Lewis, W.**
1945. Personal Communication.
46. **McMaster, P. D., & S. S. Hudack**
1935. J. Exp. Med. **61**: 783.
47. **Medawar, J.**
1940. Brit. J. Exp. Path. **21**: 205.
48. **Menkin, V., & J. Freund**
1929. Arch. Path. **8**: 263.
49. **Minot, G. R., & R. Isaacs**
1925. J. A. M. A. **84**: 1713.
50. **Mirsky, A. E.**
1943. Advances in Enzymology **3**: 1.
51. **Mudd, S.**
1932. J. Immunol. **23**: 423.
52. **Murphy, J. B. & E. Sturm**
1925. J. Exp. Med. **41**: 245.
53. **Nii, T.**
1932. Arb. 3. Abt. Anat. Inst. k. Univ. Kyoto (Series D) **2**: 70.
54. **Oesterlind, G.**
1938. Acta Path. et micro. Scand. Suppl. **34**.
55. **Ohno, R.**
1930. Bioch. Ztschr. **218**: 206.
56. **Okaue, Y., & G. Hojo**
1935–1936. Arb. 3. Abt. Anat. Inst. k. Univ. Kyoto (Series D) **5**: 62.
57. **Ono, K.**
1940. Trans. Jap. Path. Soc. **30**: 1.
58. **Ono, K., & T. Miyazaki**
1936. Trans. Jap. Path. Soc. **26**: 278.
59. **Pauling, L., D. H. Campbell, & D. Pressman**
1943. Physiol. Rev. **23**: 203.
60. **Rich, A. R.**
1936. Arch. Path. **22**: 228.
61. **Rous, J. P.**
1908. Exp. Med. **10**: 238.
62. **Sanders, A. G., H. W. Florey, & J. M. Barnes**
1940. Brit. J. Exp. Path. **21**: 254.
63. **Schumacher, S. V.**
1899. Arch. Mikr. Anat. **54**: 311.
64. **Sevag, M. G.**
1945. Immuno-Catalysis. Thomas. Springfield, Ill.

65. **Sjoevall, H.**
1936. Experimentelle Untersuchungen ueber das Blut und die blutbildenden Organe—besonders das lymphatische Gewebe—des Kaninchens bei wiederholten Aderlaessen. Hakan Ohlssons Boktryckeri. Lund.
66. **Sternberg, C.**
1926. Handb. spez. Path. Anat. & Hist. **1**(1): 249. Henke & Lubarsch.
67. **Tuta, J. A.**
1937. Fol. haem. **57**: 122.
68. **White, A., & T. F. Dougherty**
1945. Endocrinology **36**: 207.
69. **Yoffey, J. M.**
1933. J. Anat. **67**: 250.
70. **Yoffey, J. M.**
1936. J. Anat. **70**: 507.
71. **Yoffey, J. M., & C. K. Drinker**
1939. Anat. Rec. **73**: 417.
72. **Yoffey, J. M., & J. Parnell**
1944. J. Anat. **78**: 109.

DISCUSSION OF THE PAPER

Dr. J. Furth (*Cornell University Medical College, Department of Pathology, N.Y.*):

The following considerations argue against the idea of Heiberg, which postulates that aged or injured lymphocytes return to the "germinal-centers" (secondary follicles), to be destroyed there, and that their constituents are utilized for the construction of new lymphocytes. Lymphocytopoiesis occurs in absence of such centers. Macrophages, which are supposed to perform this function, are few in secondary follicles of healthy animals and are more abundant at other sites of the "reticuloendothelial system." Recent observations indicate the essential identity of macrophages at different sites. Doubtless, they phagocytize injured lymphocytes; this is readily seen in sections of spleen and lymph nodes, following irradiation with small doses of x-rays, but lymphocyte-laden macrophages are far more numerous, in other parts of the "reticuloendothelial system," than in the secondary follicles.

The hypothesis that the lymphocytes are the source of antibodies and perhaps of globulins, as well, is attractive, but the localization of this function in the "germ centers" lacks conclusive evidence.

Dr. Thomas F. Dougherty (*Department of Anatomy, Yale University School of Medicine, New Haven, Connecticut*):

The number of lymphocytes in the blood stream is determined by the number of these cells entering and leaving the blood, during a unit period of time. Thus, the number of circulating lymphocytes could decrease, if there were a greater peripheral removal, or if the potential store of these cells were decreased in the organs from which lymphocytes enter the circulation.

The "lymphopenic effect" of certain adrenal cortical hormones has been shown to be due, primarily, to the dissolution of lymphocytes in the various lymphoid tissues. This results in the depletion of stores of lymphocytes and diminishes the number available for delivery to the circulation. However, the rapidly developing lymphopenia, following adrenal cortical hormone administration, indicates that from one third to half of the lymphocytes disappear from the blood, during the first hour following treatment.

There is little evidence of nuclear disintegration among the circulating lymphocytes of the hormone-treated animals. There are some morphological alterations, which, however, cannot, at present, be taken as evidence of either degeneration

or necrosis of lymphocytes. At this stage of investigation, it does not seem likely that direct destruction of circulating lymphocytes can completely account for the lymphopenic effect of adrenal cortical secretion.

It has also been found that there is no accumulation and disintegration of lymphocytes in the large capillary beds which could account for the disappearance of these cells from the circulation. There are no indications that there is an accelerated passage of lymphocytes through the gastrointestinal mucosa, although many lymphocytes undergo dissolution in this location.

The disappearance of lymphocytes from the circulation during the adrenal cortical lymphopenia could possibly be due to the emigration of these cells from the blood into the tissues, followed by their dissolution. Thus, lymphocytes which, under ordinary circumstances, could migrate through tissues and eventually be delivered back to the blood, would be removed permanently from the circulation. The possibility that the germinal center is a major site of lymphocyte removal is indicated by the fact that it is the first area in which nuclear disintegration occurs, following adrenal cortical stimulation. However, no definite evidence of migration of lymphocytes into germinal centers has been observed. Lymphocytes do disintegrate in the germinal centers, under normal conditions, and the rate of this disintegration is stimulated in adrenal cortical-hormone-treated animals. In any case, therefore, the amount of dissolution of lymphocytes in secondary nodules would probably affect the numbers of circulating cells.

It must be emphasised that the histological technics which have been brought to bear on the question of the destination of circulating lymphocytes are insufficient to provide a conclusive answer. Other approaches to this problem are being sought at the present time.

Dr. Menkin:

It would be of interest to study the capacity of patients with hyperthyroidism to produce antibodies. Kocher, in 1908, pointed out the relative lymphocytosis existing in cases of Grave's Disease. The enhanced metabolism *per se*, in this disease, could perhaps be controlled, in other conditions, to obviate this metabolic factor as the essential element concerned in antibody production.

Dr. Michael Heidelberger (*College of Physicians and Surgeons, New York, N. Y.*)**:**

While Dr. Ehrich has presented an impressive array of experiments and data in favor of the lymphocytic formation of antibodies, it would seem that this function can still not be excluded for the cells of the reticuloendothelial system. Perhaps Dr. Ehrich does not wish to exclude them. A difficult point in the lymphocytic theory is the role assigned to the phagocytic cells: namely, of preparing the antigen for the lymphocytes. If this function should involve enzymatic activity, this could not proceed far, without destroying the antigenic properties of a protein, as Landsteiner has shown.

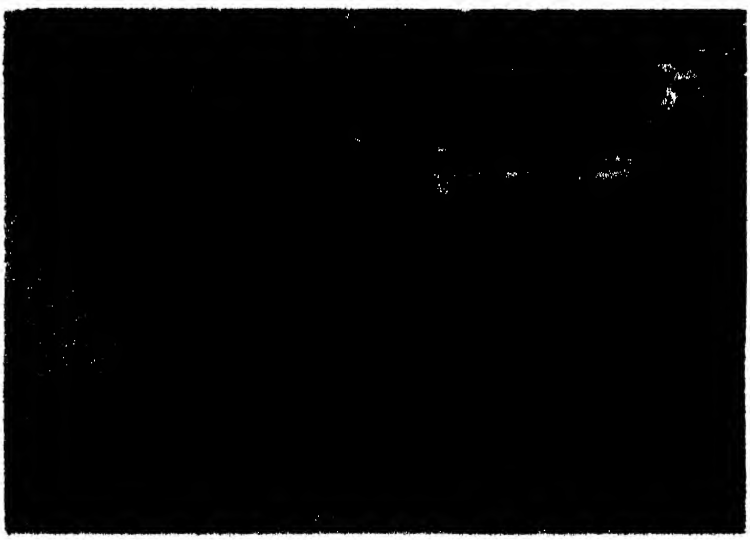
Dr. T. N. Harris (*University of Pennsylvania, Philadelphia, Pennsylvania*)**:**

The observations which indicate that the lymphocyte has a function in the formation of antibodies have led to further studies, in an attempt to define this function. A study has, quite recently, been begun of the sequence of events following the injection of an antigen and prior to the earliest appearance of antibodies. Sheep erythrocytes were injected into the pad of the rabbit's foot, and, on the first and second day following the injection, lymph was collected from the efferent lymphatic of the regional lymph node. The clear lymph plasma was found to contain material which showed immunologic similarity to the sheep erythrocyte by the inhibition of specific hemolysis. It was felt that the question of interfering antibodies, such as the Forssman, was quite probably eliminated by the high titer of the rabbit's anti-sheep erythrocyte serum used in the test (1:1800). The possibility of particles released from laked erythrocytes, quite unspecifically, and carried in the lymph stream must also be borne in mind, although quantitative considerations render this unlikely.

Repetitions of this work, which are now in progress, will, it is hoped, extend the field of observations and also tend to obviate conflicting interpretations of the experimental results with erythrocytes.

The question of the relation of the dissolved material to the whole cell injected is of considerable interest. According to the concept which has arisen in the course of these studies, this material might be the product of the action of phagocytes on the cells injected. We do not, as yet, know the size of these antigen-specific particles in the lymph. It is planned for the near future to obtain an estimate of their size, and to compare this with the size of soluble antigen-specific particles that may be derived from the particulate antigen by mechanical means, such as grinding or sonic vibration.

If such studies indicate a relatively small molecule, there does not appear, necessarily, to be any discrepancy with earlier immunologic studies, which showed that, upon enzymatic degradation of antigenic proteins, or other chemical manipulation, there is a decrease or loss of antigenic activity. It is quite possible that the cleavage of large molecules or particles within the phagocytic cells, if this occurs in the physiologic preparation of antigenic material for antibody production, is such as to spare the configuration or groups responsible for the antigenicity and specificity. This might well not be the case in the *in vitro* experiments of tryptic digestion, etc., to which reference has been made.



WILLIAM EHRICH: ROLE OF THE LYMPHOCYTE

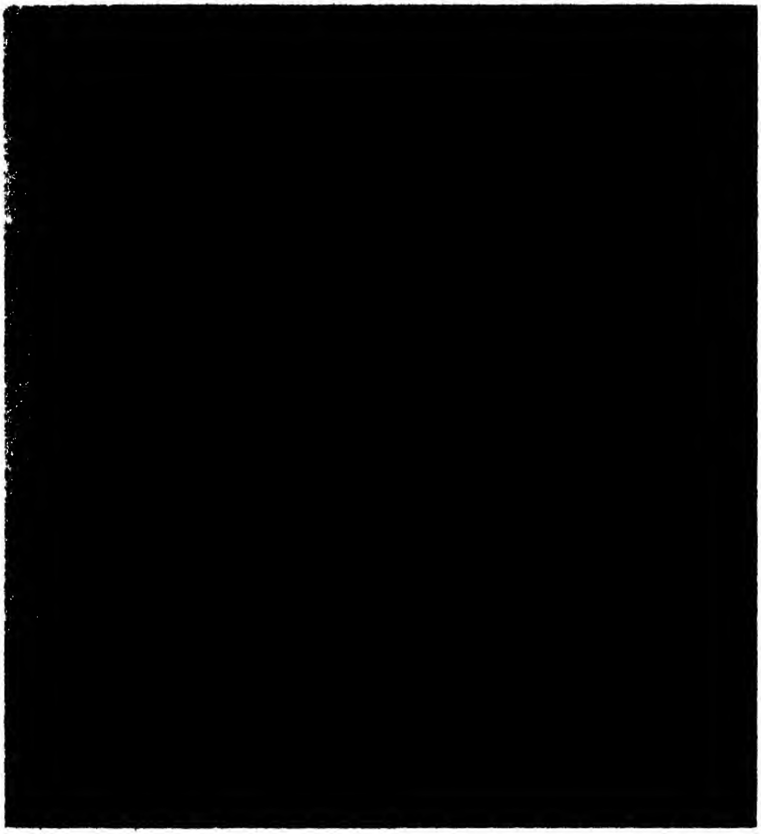
PLATE 5

Primary nodules. left x 75 (from Ehrich[15]), right x 600 (from Ehrich[15]) (Left. taken from lymph node of a rabbit; right, from node of a cat.)

PLATE 6

Secondary nodules: left x 70 (from Ehrich[illegible]); right x 500 Note the fragments of lymphocytes in the macrophages (A). (Left, taken from lymph node of a rabbit; right, from a human node.)

WILLIAM EHRICH ROLE OF THE LYMPHOCYTE



WILLIAM EHRICH: ROLE OF THE LYMPHOCYTE

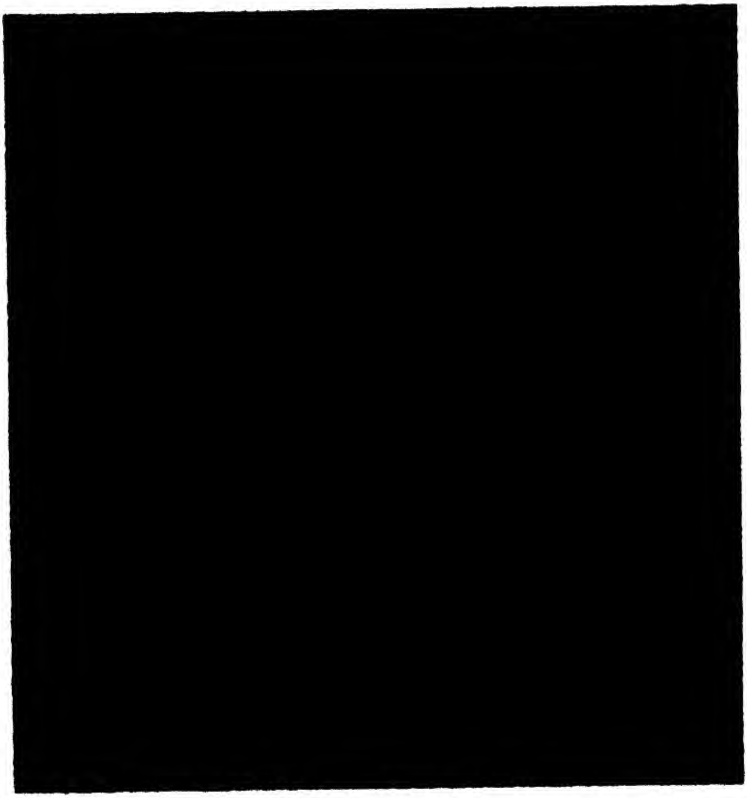
PLATE 7

Tertiary nodules, x 70. Note the primary and secondary nodules at the periphery of the tertiary nodules. (Taken from lymph node of a rabbit.)

PLATE 10

Post-capillary veins of tertiary nodules showing large numbers of lymphocytes in their lumina: the right shows artery (A) and vein (B) in the same section. Left x 600 (from Ehrich[16]), right x 600 (from Ehrich[17]). Left, taken from lymph node of a rabbit; right, from node of a cat.)

WILLIAM EHRICH: ROLE OF THE LYMPHOCYTE



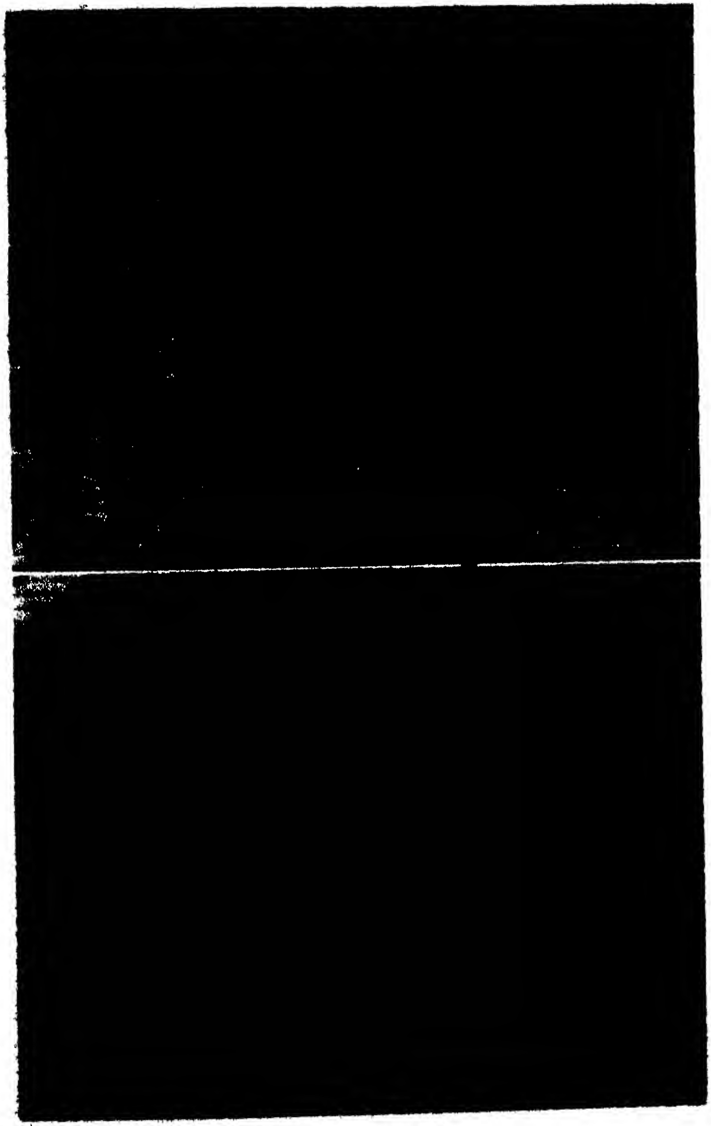
WILLIAM EHRICH. ROLE OF THE LYMPHOCYTE

PLATE 11

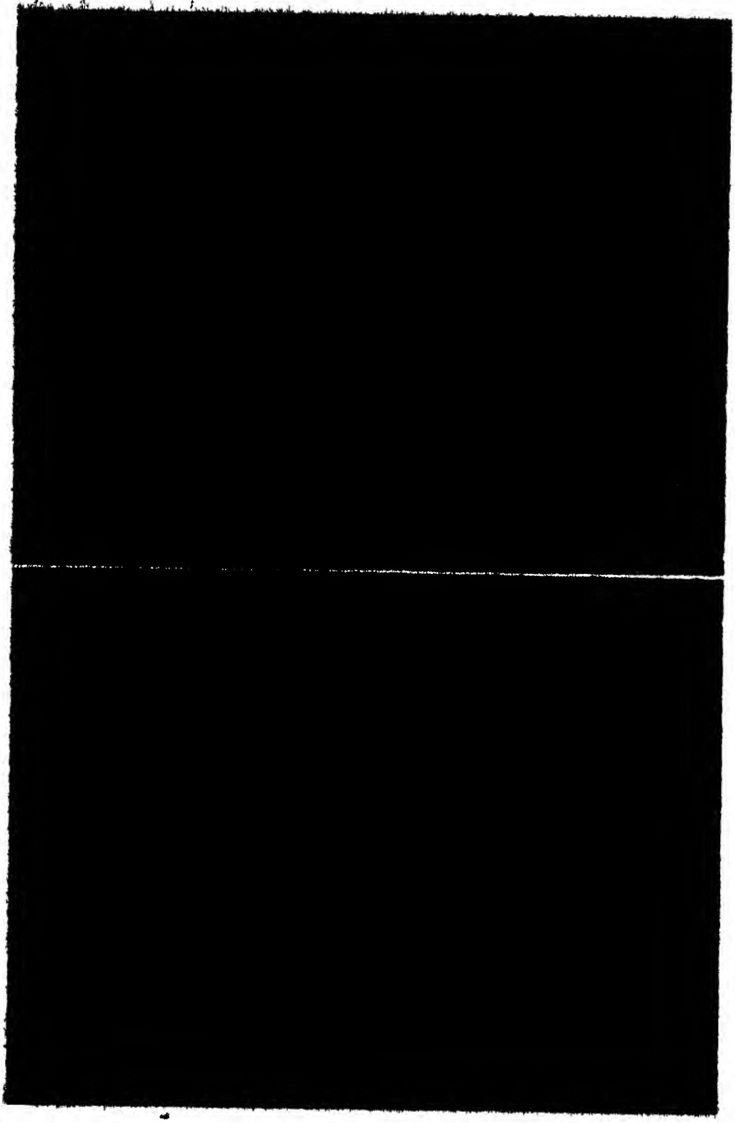
Transitional stage from secondary nodule to tertiary nodule, x 185 (from Ehrich[18]). (Taken from lymph node of a rabbit.)

PLATE 12

Sinuses filled with small lymphocytes (A) in a recently formed tertiary nodule, 3 days after injection of typhoid vaccine: left x 80, right x 600 The right is a high power of a portion of the left. (Both taken from lymph nodes of a rabbit)



WILLIAM EHRICH: ROLE OF THE LYMPHOCYTE



[illegible]: ROLE OF THE LYMPHOCYTE

Plate 13

Marked lymphocytopoiesis, left 4 days, right 6 days after injection of dysentery vaccine, x 370. Note the marked basophilia of the cytoplasm of the lymphocytes, especially after 4 days. (Both taken from lymph nodes of a rabbit.)

THE ROLE OF LYMPHOCYTES IN NORMAL AND IMMUNE GLOBULIN PRODUCTION, AND THE MODE OF RELEASE OF GLOBULIN FROM LYMPHOCYTES*

By Abraham White and Thomas F. Dougherty

Departments of Physiological Chemistry and Anatomy, Yale University, New Haven, Connecticut

INTRODUCTION

The inverse relationship between the degree of adrenal cortical secretion and thymic size[1-7] led to studies to determine whether this relationship extended to other lymphoid structures. The secretion of adrenal cortical hormones controlled by the pituitary adrenotrophic hormone is the normal mechanism regulating lymphoid tissue mass.[8, 9] Further, at a time when lymphoid tissue involution is maximal, as a result of augmented pituitary-adrenal cortical secretion, a profound absolute lymphopenia is present.[10] The diminished lymphoid tissue mass, occurring concomitantly with blood lymphopenia, posed the problem of explaining this apparently paradoxical phenomenon. In other words, it was necessary to explain why lymphocytes disappeared from both lymphoid organs and from the blood at approximately the same time.

An answer to this problem was sought in detailed histological study of lymphoid structures, at a time when lymphoid tissue involution and blood lymphopenia were most marked. Earlier studies had shown that there was no accumulation of lymphocytes in large capillary beds which would account for their disappearance from the blood. The histological studies revealed[11, 12] that the decrease in lymphoid tissue mass and the blood lymphopenia were both a result of the marked dissolution of lymphocytes which occurred in the lymphoid organs as a consequence of augmented pituitary-adrenal cortical secretion. Therefore, decreased lymphoid tissue weight was due to fewer lymphocytes in the lymphoid organs, and the lymphopenia was a result of failure of delivery or disintegration of these cells.

* The data presented in this publication were obtained in investigations aided by grants from the Josiah Macy, Jr., Foundation and the Fluid Research Fund, Yale University School of Medicine.

Dissolution of lymphocytes is a term which we have used to describe processes by which lymphocyte cytoplasm is liberated from these cells. The enhanced dissolution of lymphocytes produced by adrenal cortical hormones would be a means by which constituents of lymphocytes could be released in large amounts to the lymph, and thus to the blood.

Since proteins are important components of all cells, the kind and quantity of protein arising from lymphocyte dissolution became of interest. It appeared possible that, of the unknown functions of the lymphocyte, one might be a contribution of globulin to the serum, particularly since the globulins of the blood have long been believed to be of extra-hepatic origin. Moreover, a particular globulin, the γ-globulin of the serum, has been demonstrated to carry most of the antibodies, and lymphoid structures have been recognized as a probable site of antibody production.[13-16]

This paper will be concerned with the presentation of data which establish the following:

1. One of the normal fates of lymphocytes is their continual dissolution.

2. The rate of dissolution of lymphocytes is regulated by pituitary-adrenal cortical secretion.

3. The normal dissolution of lymphocytes liberates from these cells a protein which is identical with γ-globulin.

4. In the immunized animal, antibodies are present in lymphocytes; dissolution of the cells, therefore, liberates immune globulin.

5. The mode of release of globulin from lymphocytes is normally effected by a shedding or budding of lymphocyte cytoplasm.

6. The enhancement of normal or immune globulin in the blood, as a consequence of exposure to a wide variety of physiological and pathological conditions, is a result of stimulation of the pituitary-adrenal cortical secretory mechanism.

EXPERIMENTAL OBSERVATIONS

The daily injection of pituitary adrenotrophic hormone in mice, for 15 days, produced a significant decrease in the total mass of lymphoid tissue.[8] On the other hand, adrenalectomy is followed by lymphoid tissue hypertrophy.[17, 18] Therefore, it was apparent that adrenal cortical steroids are normally concerned with the regulation of lymphoid tissue mass. These steroid hormones might be exerting their influence as a result of accelerating the rate of removal of lymphoid cells, or by inhibiting the production of these cells.

Adrenal Cholesterol as an Index of Adrenal Cortical Secretion

In initiating this work, it was necessary to determine when the maximum secretion of adrenal cortical steroids occurred following a single injection of adrenotrophic hormone. As an index of the amount of adrenal cortical activity, adrenal cholesterol values were determined in the mouse, at varying intervals, following single injections of the hormone.[10] It is well known that depletion of adrenal lipids occurs at a time when the metabolic effects of adrenal cortical steroids are being manifested.

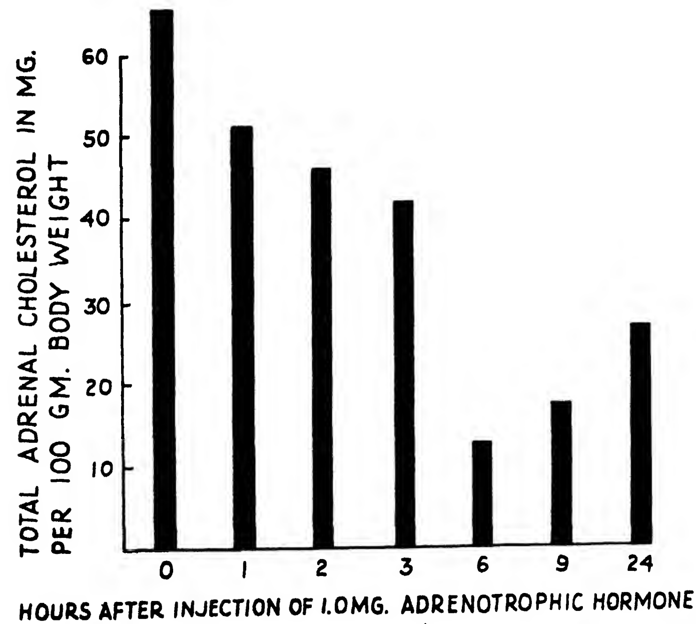


FIGURE 1. Effect of a single injection of adrenotrophic hormone on the total adrenal cholesterol of normal mice. Each value represents the average of data from paired adrenals from each of two animals.

FIGURE 1 shows the alterations occurring in total adrenal cholesterol of the mouse, following the subcutaneous injection of one milligram of adrenotrophic hormone. Maximum depletion of cholesterol occurred 6 hours after subcutaneous injection of adrenotrophin. Thereafter, the level of adrenal cholesterol began to return to normal. It had been

demonstrated previously that a single intraperitoneal injection of adrenotrophin in the rat produced maximal cholesterol depletion within 3 hours.[19, 20]

Adrenal Cortical Control of the Number of Circulating Lymphocytes

Having demonstrated that maximal adrenal cortical secretion occurs at 6 to 9 hours under these experimental conditions, further studies of the physiological effects of a single injection of adrenotrophin were conducted. In animals which were injected daily for 15 days, it had been observed that, 24 hours after the previous injection, a slight, but significant lymphopenia was present.[21] Therefore, blood studies were made at intervals shortly following adrenotrophic hormone injection.

Within 1 hour after a single injection of adrenotrophin in the mouse, the absolute number of lymphocytes declined, and a striking lymphopenia was present 9 hours following hormone administration.[10] Thereafter, the number of circulating lymphocytes increased, and, 24 hours after hormone administration, was only slightly less than that seen in normal, untreated animals. It is interesting, in this connection, that Reinhardt and Li[22] have recently reported that, within 30 minutes following a single injection of adrenotrophic hormone in the rat, there is a 50 per cent decrease in the number of lymphocytes in the lymph collected from the thoracic duct. This confirms our conclusion that the adrenal cortex controls the numbers of circulating lymphocytes.

The lymphopenic effect of adrenotrophic hormone has been demonstrated in a number of other species, *e.g.*, rat,[10] rabbit,[10] and dog.[18, 23]

The adrenotrophic hormone has no effect on the number of blood lymphocytes in adrenalectomized animals.[10] Therefore, the effect of this hormone is mediated by way of the adrenal cortex. Also, other pure proteins, such as prolactin and human serum γ-globulin, have no effect on the lymphocyte levels of the blood of normal animals.[10] Therefore, the action of the adrenotrophic hormone is due to its property as a hormone, and is not a non-specific protein effect.

The steroids of the adrenal cortex which are liberated by the adrenotrophic hormone, also, as might be expected, produce a lymphopenia in normal animals. Moreover, these steroids and whole adrenal cortical extract are capable of producing complete replacement therapy in the adrenalectomized animal. The effect of adrenal cortical extract is evident earlier than in the case of adrenotrophic hormone and does not last for as long a period of time. The lymphopenic effect which has been discussed is produced by those adrenal cortical steroids which

have an oxygen atom in position 11 of the steroid nucleus, *e.g.*, corticosterone and compound E. On the other hand, desoxycorticosterone acetate, which is not oxygenated in position 11, has no effect on the numbers of circulating lymphocytes in the normal or the adrenalectomized animal.[10]

The observation of lymphopenia, following the injection of adrenotrophic hormone or adrenal cortical extract, raised the question of its cause. Obviously, since lymphocytes arising in lymphoid organs are constantly entering the circulation *via* the thoracic duct, the lymphopenia would be due to a failure of delivery of these cells to the circulation. The reasons for this failure were sought in a detailed histological study of lymphoid structures following injection of hormone.

Effect of Adrenal Cortical Hormones on Lymphoid Tissues and Lymphocytes

Within a few hours after adrenotrophic or adrenal cortical hormone injection, the lymphoid tissues are edematous and swollen, and, in the medullary portion of the lymph nodes and thymus, the edematous fluid contains lymphocytes undergoing degeneration (PLATE 14, FIGURES 1 and 2). This edematous change is evident in mice and rabbits, following hormone injection, has been observed in all the animals studied, and occurs in all of the lymphoid tissues. The edema which appears early leads to an increase in weight of the lymphoid structures. However, within 15 hours, the edema has subsided, and, due to the loss of fluid and the presence of fewer lymphocytes, the lymphoid tissue mass actually has a smaller weight than that seen in normal, untreated animals.

The initial effects of adrenotrophic or adrenal cortical hormones on lymphocytes are seen at approximately 1 hour in all lymphoid tissues. In the thymus, degeneration of lymphocytes is first observed in the medulla, and subsequently degeneration and necrosis of these cells occurs in the cortex, so that, 3 hours after hormone injection in mice, and after 6 hours in rabbits, there is an accumulation of large amounts of nuclear debris which is being phagocytized by macrophages (PLATE 14, FIGURE 3). This produces a pitted appearance of the thymic cortex and is one of the most striking alterations seen in the lymphoid tissues following adrenal cortical stimulation, whether by adrenotrophic hormone or by non-specific toxic agents. In lymphoid organs other than the thymus, injection of hormone produces a sequence of changes which are characteristic of all the lymphoid tissues. The initial effect on

lymphocytes is seen in the germinal center, which becomes edematous and filled with degenerating lymphocytes and much nuclear debris resulting from karryorhexis (PLATE 14, FIGURE 4). A zone of degenerating lymphocytes radiates from the germinal center and encompasses the cells in the rest of the node. These changes take place in the white pulp of the spleen, in the Peyer's patches, and in the appendix of rabbits. Following this stage of edema and degeneration of lymphocytes, there occurs a reparative phase, during which macrophages phagocytize the nuclei and nuclear particles. Within 9 hours, mitotic figures may be seen, although they are not numerous at this time. Subsequently, the lymphoid tissues return to their normal appearance, although they are much smaller and do not contain normal numbers of lymphocytes.

The disappearance of lymphocytes from the lymphoid structures is due to their dissolution, and this establishes the lymphopenia as due to a failure of delivery of lymphocytes to the circulation. The term, "dissolution of lymphocytes," is used to include the various alterations in these cells which are seen following injection of adrenotrophic or adrenal cortical hormones. One of these alterations is an increased loss of cytoplasm by a budding or shedding process (PLATE 14, FIGURE 5). This phenomenon is known to occur normally in the life of the lymphocyte[24] and is greatly enhanced by hormone administration. The loss of cytoplasm may leave completely denuded lymphocyte nuclei with a normal appearance. It is not unlikely that these nuclei may regenerate cytoplasm. A second alteration in lymphocytes is a destruction of the nuclei, which occurs both by karryolysis and karryorhexis. Apparently, most of the lymphocytes undergoing necrosis show actual destruction of the nuclei (PLATE 14, FIGURE 5). A third type of change, which is present in the lymphocytes, is the development of hyaline granules in the cytoplasm which is freed from the cells and found in the lymph (PLATE 14, FIGURE 5).

The series of alterations which has been characterized by the term, "dissolution of lymphocytes," occurs as a result of the physiological action of adrenal cortical steroids. Therefore, the lymphocyte is the target cell of these steroid hormones. The histological changes discussed here have been described in detail, elsewhere.[12]

The Lymphocyte as a Source of Serum Protein

The budding of the cytoplasm of the lymphocytes, with the liberation of this material into the lymph, and subsequently into the blood, raised the problem of the possible contribution of constituents of

lymphocyte cytoplasm to the blood. Inasmuch as lymphoid structures have been suggested as a site of antibody formation, and since lymphocyte cytoplasm is obviously rich in protein, alterations in blood proteins were studied at intervals, following injection of adrenotrophic hormone or adrenal cortical steroids.

Time intervals after hormone administration were chosen to correspond with maximum blood lymphopenia and maximum degree of lymphocyte dissolution in the tissues. At 3 and at 6 hours after a single injection of adrenotrophic hormone in the rat, there is a significant increase in the level of total serum proteins (TABLE 1). Subsequently, the total serum protein concentration returns to a value approximately that of normal animals. The continued, daily injection of adrenotrophic hormone in the mouse will sustain the serum protein level at a value significantly greater than normal. Removal of the adrenals in the mouse, with the administration of desoxycorticosterone acetate, in order to maintain an approximately normal blood volume, leads to a significant decrease in the total serum protein level (TABLE 1). In the adrenalectomized animal, therefore, the removal of the mecha-

TABLE 1

EFFECT OF PITUITARY ADRENOTROPHIC HORMONE AND OF ADRENALECTOMY ON TOTAL SERUM PROTEINS

Animal	No. of animals	Treatment	Total serum proteins gm./%	P values**
Normal rat	22	......	6.00 ±0.01*	
Normal rat	8	3 hrs. after 3 mg A***	6.22 ±0.07	<0.01
Normal rat	8	6 hrs. after 3 mg A	6.30 ±0.08	<0.01
Normal rat	4	24 hrs. after 3 mg A	5.81 ±0.01	0.04
Normal mouse	19	......	6.02 ±0.02	
Normal mouse	17	1 mg A daily, 15 days	6.52 ±0.07	<0.01
Adrenalectomized mouse	16	0.025 D**** daily, 8 days	5.59 ±0.08	<0.01

* Means and Standard errors;
** P values, compared to controls;
*** A = adrenotrophic hormone;
**** D= desoxycorticosterone acetate (Schering).

nism controlling the rate of lymphocyte dissolution has decreased the rate at which protein is contributed to the blood from lymphoid structures.

A more detailed analysis of the serum protein picture following

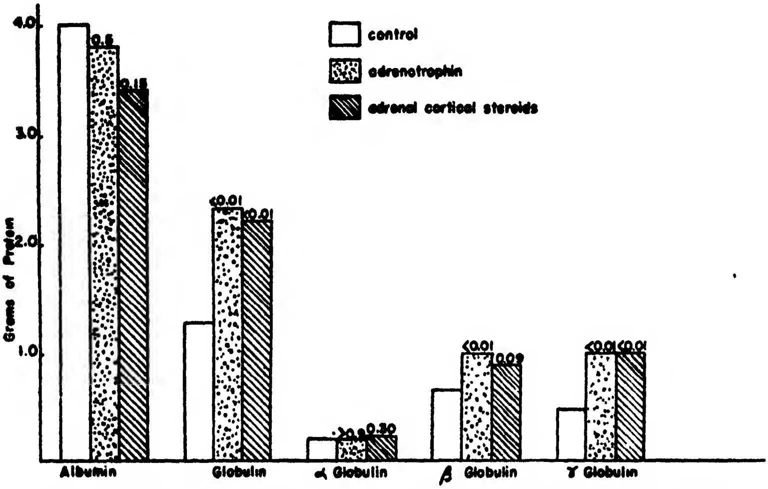


FIGURE 2. Serum protein changes in rabbits, within 24 hours after injection of adrenotrophic hormone or adrenal cortical steroids. The control series represents average values on a group of 5 animals. The adrenotrophic hormone studies were made on 7 animals; the adrenal cortical steroid injections, on 5. Only one blood sample was taken from each rabbit, and the time intervals studied were 3, 6, 9, 12, and 24 hours after hormone injection. The data at each time interval were similar and, consequently, have been averaged. The values at the top of each column are the statistical "P" values, as compared to the controls.

hormone administration was obtained by the use of the Tiselius apparatus.[25] The experiments consisted of a study of the blood of rabbits, following a single injection of either adrenotrophic hormone or adrenal cortical steroids. The animals were examined at intervals of from 3 to 24 hours following treatment with hormone, and the blood from the animals was examined electrophoretically, in order to determine the serum protein pattern.

The data which have been obtained are shown in FIGURE 2. It will be seen that the injection of adrenotrophic hormone or adrenal cortical steroids produces a lowering in the total albumin concentration of the serum, which, however, is not significant. The total globulin concentration is markedly increased following administration of adrenotrophic hormone or adrenal cortical steroids. Examination of the electrophoretic pattern reveals that there is no alteration in the alpha globulin fraction, but that both the β- and γ-globulin fractions are markedly elevated.

The time correlation existing among blood lymphopenia, lymphocyte dissolution, and serum beta and gamma globulin increases, following hormone injection, strongly suggested that lymphocytes contain globulin, which is being released to the blood in circumstances of augmented pituitary adrenal cortical secretion. Consequently, an attempt was made to identify the protein components of lymphocytes

Protein Constituents of Lymphocytes

Lymphocyte extracts were prepared by mincing lymphoid tissue, suspending the mince in physiological saline, centrifuging and re-sus-

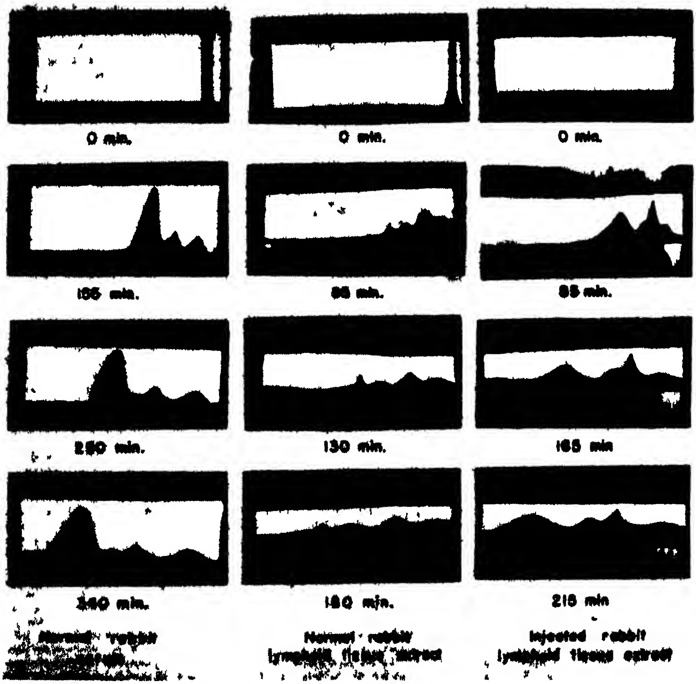


FIGURE 3. Electrophoretic patterns of normal rabbit serum and of lymphoid tissue extracts All photographs of descending limb at times indicated. Temp. 3° C, pH 7 95, 0 02 molar phosphate buffer containing 0 1 molar sodium chloride. Potential-gradient 5 to 6 volts cm $^{-1}$.

pending several times and centrifuging, in order to wash the cells free of lymph. The stained smears of the sediment so obtained showed that approximately 90 per cent of the cells were intact, normal lymphocytes.

These lymphocytes were then lysed, by grinding with one volume of water, and extracted with an equal volume of 10 per cent salt solution. The protein-containing extracts obtained in this manner were dialyzed against several changes of a suitable buffer and examined in the Tiselius apparatus. A characteristic series of patterns obtained in such studies is shown in FIGURE 3, which also contains the pattern of normal rabbit serum for purpose of comparison. It will be seen that there are at least four protein components present in lymphocyte extracts and that the number of these components is the same, whether one examines lymphoid tissue from normal animals, or from rabbits which have been injected with adrenotrophic hormone or adrenal cortical steroids.

However, in the extracts of lymphocytes obtained from hormone-treated rabbits, the most rapidly moving component is present in greatly augmented amounts. It is probable that this component is a nucleoprotein. The slowest component in lymphocytes has an electrophoretic mobility which is identical with that of the normal γ-globulin of the serum. Another component of lymphocyte extracts appears to be identical with the β-globulin fraction of rabbit serum.*

The demonstration of normal γ-globulin in lymphocytes suggested that labeled globulin or antibodies might be present in the lymphocytes of immunized animals. Mice were immunized to sheep erythrocytes, and at the time that these animals had demonstrable circulating antibody, they were sacrificed, and titer estimations were made on the blood and on extracts of lymphocytes prepared as previously described.

Some of the data which have been obtained[26] are shown in TABLE 2. It will be seen that the lymphocytes of non-immunized mice contain no antibody. In contrast, lymphocyte extracts from immunized mice contained a significant amount of antibody. The serum of the immunized animals, of course, also contains antibody, but it should be noted that the titer in serum is less than that observed in the lymphocyte extracts, despite the fact that the serum contained approximately three to four times as much nitrogen as that present in the lymphocyte extracts.

---

* Electrophoretic studies have also been made in barbiturate buffer at pH 8.6. The preliminary data indicate that, in this buffer, lymphocyte extracts show three to four components, depending on the total protein content of the extract being examined. It may also be added that, in unpublished studies, it has been possible to prepare extracts of lymphocytes having a higher protein concentration than previously obtained.[25] When studied electrophorectically in phosphate buffer of pH 7.25, these extracts appear to have at least five protein components, as compared to the four reported in this paper and previously described.[25] In both phosphate and barbiturate buffers, regardless of the number of components seen in the electrophoretic experiments, one of these components has a mobility identical with that of normal gamma globulin of the serum, under [illegible]ntal conditions.

TABLE 2

AGGLUTININ AND HEMOLYSIN TITERS IN SERA AND IN TISSUE EXTRACTS OF NORMAL AND OF IMMUNIZED MICE

Material titrated	Mg. nitrogen/ml. sera or extract			Agglutinin titers			Hemolysin titers		
	Normal mice	Immunized mice I	Immunized mice II	Normal mice	Immunized mice I	Immunized mice II	Normal mice	Immunized mice I	Immunized mice II
Lymphocyte extract	2.26	2.00	2.10	0*	1-2560	1-2560	0	1-3000	1-3000
Serum	7.70	7.39	8.21	0	1-1280	1-1280	0	1-2000	1-2000
First lymphocyte washings	—†	—	—	0	0	0	0	1-10	1-10
Second lymphocyte washings	—	—	—	0	0	0	0	0	0
Third lymphocyte washings	—	—	—	0	0	0	0	0	0
Salivary gland extract	—	—	3.22	—	0	0	—	0	0
Muscle extract	—	—	2.37		0	0	—	0	0

* 0 indicates absence of titer in any dilution.
† — indicates determination not made.

Therefore, on the basis of nitrogen content, lymphocytes contained from six to eight times as high a concentration of antibody as was present in the serum of the same animal.

It will also be seen that the lymphocyte washings contained no agglutinin titer. The small amount of hemolysin titer present in the first washing was undoubtedly due to both adhering lymph and the fact that some cells were broken in the mincing process. Salivary gland extracts were of interest, because these extracts were made from a tissue containing a high proportion of cells and protein. Muscle extracts were examined, as being characteristic of connective tissue, and because of the high globulin concentration in muscle. Liver was not studied, because it is difficult to separate completely the blood from the liver tissue. It is evident from the data that no antibody is present in the salivary gland and muscle. Thus, the globulins of these tissues, even of immunized mice, are not antibody globulin. This further emphasizes the fact that the immunized animals contain a globulin which is peculiarly characteristic of the lymphocyte, namely, antibody globulin.

In other experiments,[27] anti-hemolysins to staphylococcus toxin have been demonstrated in the washed lymphocytes of mice immunized to the toxin. These observations have been made in both immunized animals and in previously immunized animals having no circulatory antibody. Malignant lymphocytes have also been demonstrated to contain antibody.[27]

Pituitary-Adrenal Cortical Control of Antibody Release from Lymphocytes

The demonstration of labeled globulin or antibody in lymphocytes led to a study of the release of this globulin by adrenotrophic hormone or adrenal cortical extracts. It will be recalled that the release of normal γ-globulin had already been demonstrated in previously discussed studies.

Two types of experiments were conducted:[28] In the first, rabbits were immunized to sheep erythrocytes, and, after the appearance of a significant quantity of circulating antibody in the blood, were permitted to remain in the laboratory without further antigen injection. After a period of approximately three months, when no circulating antibody was present, the animals were divided into groups and treated in the following manner:

In one group, each animal was given a single injection of adrenal

cortical steroids in oil; in a second group, a single injection of adrenotrophic hormone; a third, a single injection of aqueous adrenal cortical extract; and a fourth, a single injection of the original antigen, namely, sheep cells. It will be seen from the data in FIGURE 4 that injection of adrenotrophic hormone or adrenal cortical extracts produced a marked release of antibody to the blood, at a time when it has been dem-

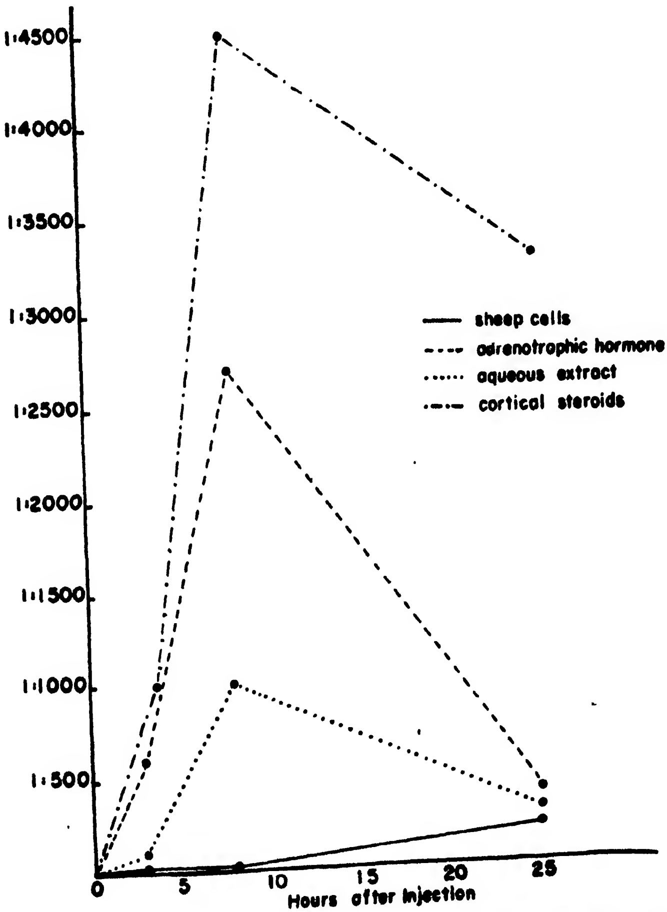


FIGURE 4. Anamnestic response in rabbits. Each curve is the average data for a group of 8 rabbits.

onstrated that there is maximal dissolution of lymphocytes in the lymphoid structures and the most profound blood lymphopenia. The release of antibodies produces a titer in the blood which is almost as great as that seen in the animals at the time when they were hyperimmunized. Following the single injection of hormone, the titers gradually approached their original level.

Another group of experiments was performed with mice also immunized to sheep erythrocytes. When these animals had a blood titer of approximately 1 to 640, immunization was stopped and they were permitted to remain in the laboratory until no antibody was present in the blood. At this time, several types of experiments were made, and the experimental procedures and data are given in FIGURE 5.

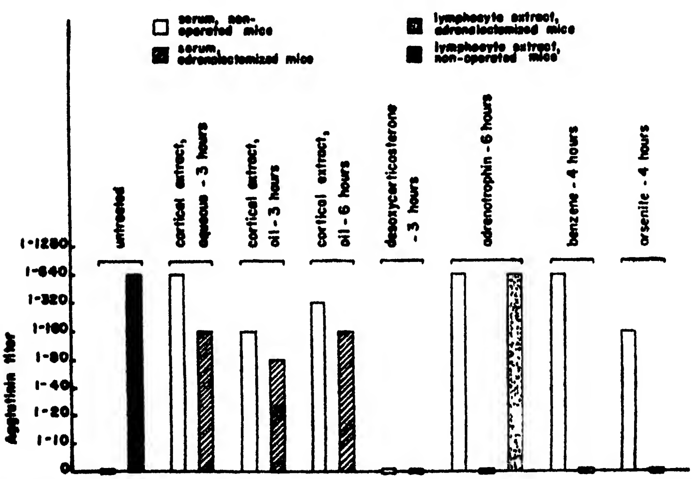


FIGURE 5. Anamnestic response in mice. Each experiment represents the pooled sera and aggregate lymphoid tissue of 5 animals.

The serum of the animals, as was indicated, contained no antibody. However, lymphocyte extracts of these animals had significant quantities of antibody. A single subcutaneous injection of aqueous adrenal cortical extract in the normal animals produced, within 3 hours, a release of antibody from the lymphocytes into the serum. Similarly, in adrenalectomized mice, adrenal cortical extract therapy was effective in producing this release of antibody. Another adrenal cortical preparation, adrenal cortical steroids in oil, was also effective in producing a release of antibody from lymphocytes in both normal and adrenalectomized animals, 3 and 6 hours following administration of the hor-

mone preparation. In contrast to the active adrenal cortical steroids, desoxycorticosterone acetate, which affects neither the numbers of blood lymphocytes, nor lymphoid tissue histology, in the dose used, does not produce a release of antibody from lymphocytes in normal, nor in adrenalectomized, animals.

Adrenotrophic hormone, by virtue of its stimulation of the adrenal cortex, also produces antibody release from lymphocytes in the normal mouse. However, in the absence of the adrenals, the pituitary hormone has no effect. It is interesting to note that, in the animals which were adrenalectomized and in which adrenotrophic hormone had no effect, the lymphocytes contained significant quantities of antibody which could not be released in the absence of adrenal cortical steroids. Two stimuli which are known to produce lymphocyte dissolution, namely, benzene and potassium arsenite, have been studied in detail with respect to their effects on lymphoid tissue histology, blood lymphocyte levels, serum protein levels, and release of antibody in normal and adrenalectomized animals. It may be said that these two agents have no effect on lymphoid tissue histology, blood lymphocytes, or serum protein levels in the adrenalectomized animal, whereas, in the normal animal, each of these two agents produces all the effects on lymphoid tissue physiology which are seen with adrenotrophic hormone or adrenal cortical extracts.[18]

The data in FIGURE 5 show one of these effects. In the normal animal, benzene and potassium arsenite produced a release of antibody from lymphocytes, but, in the adrenalectomized animal, these two agents were entirely without effect.

A crucial test of the hypothesis that adrenal cortical steroids produce these manifestations of lymphoid tissue function and histology, by virtue of their capacity to produce lymphocyte dissolution, was sought with the use of an agent which, in small doses, would produce lymphocyte dissolution only by pituitary-adrenal cortical stimulation, but, in large doses, could cause a direct destruction of lymphocyte elements, without hormone mediation. Such an agent was found in X-rays.[20] After many experiments on several hundred mice, a dose of X-radiation was found which affected lymphoid tissue structure and function only in the presence of the adrenals. Another considerably larger dose was found which produced alterations in lymphoid tissue in the presence or the absence of the adrenals. With the use of these two doses of radiation, it was then possible to study in detail the manifestations of lymphoid tissue structure and physiology which have pre-

viously been discussed in relation to adrenotrophic hormone and adrenal cortical extracts. Depletion of adrenal sudanophilic material, which is taken as an index of adrenal cortical secretion, was produced with 10 r, in the normal animal, and was also evident, to a striking degree, with 200 r. Dissolution of lymphocytes was effected by 10 r, in the normal animal, but was not produced with the same radiation dose in adrenalectomized mice. 200 r, on the other hand, produced lymphocyte dissolution in the presence or in the absence of the adrenals, probably because of its direct effect in destroying lymphocytes. As a result of the lymphocyte dissolution, there was a lymphopenia following 10 r in the normal animal, but no change in circulating lymphocyte numbers following the same dose of radiation in adrenalectomized mice. 200 r produced a lymphopenia in the absence of the adrenals. The release of protein, as a consequence of lymphocyte dissolution, was manifested by an increase in γ-globulin in normal mice irradiated with 10 r, whereas the same radiation dose in operated animals did not produce a significant alteration in the γ-globulin concentration of the blood. 200 r in either normal or adrenalectomized mice produced a striking rise in the blood γ-globulin.

Antibody globulin was released in normal, previously immunized mice, following 10 r, but was not evident in the blood of adrenalectomized mice given the same radiation dose. On the other hand, 200 r was clearly effective in producing a release of antibody from the lymphocytes of adrenalectomized animals. These data are supporting proof of the hypothesis that the action of the adrenal cortex in producing serum protein and antibody increases is based on the dissolution of lymphocytes by the steroid hormones of this gland.

DISCUSSION

The control of lymphoid tissue structure and function by pituitary-adrenal cortical secretion is evident from the alterations seen in this tissue following pituitary stimulation or administration of excess adrenotrophin or adrenal cortical preparations. These alterations are: a dissolution of lymphocytes in lymphoid structures;[11, 12] a profound lymphopenia;[10] an increase in serum β- and γ-globulins;[25] and, in the immunized animal, a release of antibody globulin to the circulation.[28, 30, 30a] The dissolution of lymphocytes is characterized by several types of alterations in these cells. One of the most prominent of these is a shedding or budding of the lymphocyte cytoplasm. As a consequence, this material is released to the lymph and, thus, to the systemic

circulation. Lymphocytes, in the normal animal, have been demonstrated to contain a protein identical with the normal serum γ-globulin;[25] a second component of lymphocytes is probably identical with the β-globulin of serum. Therefore, the accentuated cytoplasmic budding of lymphocytes, as a consequence of augmented adrenal cortical hormone concentration, results in an increased rate of contribution of lymphocyte globulins to the blood. In the immunized animal, lymphocytes contain antibody,[26] and the release of these labeled globulins produces a rise in the level of circulating antibody.[28, 30, 30a]

The increased dissolution rate of lymphocytes within lymphoid structures results in a failure of delivery of these cells to the circulation. It has been conclusively demonstrated that the adrenal cortical steroids controlling lymphocyte dissolution are those which are oxygenated in position 11 of the steroid nucleus.[11, 12] Desoxycorticosterone acetate is without effect on lymphoid tissue structure and function. Thus, the peripheral lymphopenia following acute adrenal cortical stimulation, whether by administration of adrenotrophic hormone or by activation of the pituitary-adrenal cortical mechanism by a variety of non-specific stimuli, is a readily determined manifestation of adrenal cortical activity. The presence of an acute lymphopenia may be accepted as evidence that the entire sequence of events described previously is taking place.

The demonstration that lymphocytes are a potential source of serum globulins emphasizes the importance of lymphoid tissue in protein metabolism. Lymphocytes are the most widely distributed cells in the body. If the total number of lymphocytes were aggregated, they would form a rather large organ. Since the adrenal cortical steroid hormones have been demonstrated to produce dissolution of single lymphocytes, as well as aggregates of these cells in the lymphoid organs, it is apparent that it is the number of lymphocytes, rather than the aggregation of these cells, which is the significant factor relating to the contributions which lymphocytes make to protein metabolism. The release of the protein of lymphocytes under circumstances of augmented pituitary-adrenal cortical secretion suggests that lymphocytes are a storehouse of readily available protein. This protein may constitute a significant portion of the reserve or deposit protein which investigators have discussed for many years, and would be the source of a large proportion of the extra-hepatic nitrogen which appears in the urine in conditions of stress.[31] It has recently been demonstrated[32] that lymphoid tissue involution and the loss of nitrogen from this tissue

as a result of fasting will occur only in the presence of intact adrenals. In the adrenalectomized animal, the reserve protein present in lymphocytes cannot be released during a period of fast. The availability of this protein in other circumstances of stress is also probably dependent upon the presence of the pituitary and the adrenals.

Whipple and his colleagues[33, 34] have postulated the existence of a dynamic equilibrium between plasma and tissue proteins. The intimate relationship between plasma globulins and lymphocyte proteins has been demonstrated in the present investigations. The cytoplasmic proteins of lymphocytes and the hormonal control of their release, therefore, play a prime role in studies of plasma protein regeneration. This attains added significance in the light of the variety of stimuli, including hemorrhage, chronic inanition, injection of heterologous protein or protein derivatives, and irritation of the peritoneum, each of which is a potent activator of the pituitary-adrenal cortical mechanism.[20]

The demonstration of the high nutritive value of γ-globulin,[35] which has now been shown to be a normal constituent of lymphocytes, gives added importance to the role of these cells in protein metabolism. Indeed, the lymphocyte may bear a relationship to protein metabolism similar to that of the fat cell to lipid metabolism.

In the immunized animal, a portion of the protein of lymphocytes is labeled globulin or antibody. The rate of release of this protein is regulated by the same factors controlling protein release from lymphocytes in the non-immunized animal. One manifestation of the pituitary-adrenal cortical control of antibody release from lymphocytes is the anamnestic response.[28] It is interesting, in this connection, to point out that Cannon[36] suggested that the increased labeled protein appearing in the serum during the anamnestic reaction must arise from reserve protein, and that an explanation of the anamnestic response would contribute to identification of this reserve or stored protein. The demonstration of the basis of the anamnestic response is proof that lymphocytes are a storehouse of protein. Animals with no circulating antibody have been shown to have antibody within their lymphocytes, and this labeled globulin is released to the circulation, as a result of increased supply of pituitary-adrenal cortical hormones.[28] This increase may occur as a result of administration of adrenotrophin or adrenal cortical steroids, in the normal animal, and of adrenal cortical extracts, in the adrenalectomized animal. Furthermore, increased secretion of the animal's own adrenotrophic hormone may occur in re-

sponse to a wide variety of unrelated stimuli which are known to activate pituitary-adrenal cortical secretion. Thus, hemorrhage, cold, heat, bacterial toxins, foreign proteins, and toxic chemical agents, are all agents which activate the hormonal mechanism[20] and have been used to demonstrate the anamnestic response.[14, 36]

Benzene, arsenite, and X-radiation have been studied in some detail in the present investigations. The doses of benzene and of arsenite employed effected a release of antibody from lymphocytes only in the presence of the adrenals. In the case of X-rays, a low dose of radiation produced an increase in γ-globulin and an anamnestic response only in unoperated mice. However, a large radiation dose effected normal and immune globulin release from lymphocytes, even in the absence of the hormonal mechanism in the adrenalectomized animals.[29] In these experiments, detailed studies of lymphoid tissue alterations revealed that lymphocyte dissolution occurs with high radiation doses, in adrenalectomized animals. In addition to the anamnestic response and lymphocyte dissolution, the operated animal also shows a lymphopenia with a large dose of X-rays. The occurrence of the characteristic sequence of events, which have been shown to be related to lymphoid tissue structure and function, in adrenalectomized mice given high radiation doses, is a result of the breaking up of lymphocytes without intervention of the normal hormonal mechanism. Thus, under normal circumstances, the control of lymphoid tissue structure and function by adrenal cortical secretion is based upon the ability of the steroid hormones to produce lymphocyte dissolution.

The conclusive demonstration that lymphocytes contain antibody leads to the question of whether these cells may form antibody. Antibody production may be considered to have two aspects. One is concerned with the site and circumstances of synthesis of the first altered globulin molecules. The other relates to factors increasing the numbers of antibody-carrying cells. The primary site of antibody production is a matter of conjecture. Several investigators have suggested that reticulo-endothelial cells are concerned with the production of immune globulin.[13, 14] Recently, Ehrich and Harris[37] have discussed the various theoretical conditions by which the lymphocyte may become a carrier of antibody.

Inasmuch as lymphocytes have been demonstrated to contain antibody, it is evident that, as these cells reproduce themselves, the labeled globulin constituent of the cytoplasm is transferred to the daughter cells. Thus, tumors developing from transplants of antibody-contain-

ing malignant lymphocytes contained as much immune globulin as the tumor from which the tissue was taken for transplantation.[27] Therefore, factors which increase numbers of lymphocytes will also augment antibody production, by increasing the available potential store of immune globulin. Under circumstances of augmented pituitary-adrenal cortical secretion, this globulin is released to the circulation. The quantity released will be directly proportional to the numbers of antibody-carrying lymphocytes in the lymphoid structures.

The presence of antibody in lymphocytes gives added significance to the wide distribution of these cells in the body. It may be suggested that at least one major function of the lymphocytes which are infiltrating into all of the tissues of the body is to serve as a medium of distribution of their important cytoplasmic protein, γ-globulin. The recent demonstration of the high nutritive value of γ-globulin suggests that lymphocytes supply to cells throughout the body a protein which may fulfil nutritive and, perhaps, other physiological needs of the body cells.

SUMMARY

1. Lymphocytes contain at least one, and probably two, globulins which are identical with globulins of the blood.

2. One of the normal fates of lymphocytes is a dissolution of these cells. The rate of this process is under pituitary-adrenal cortical control and results in a contribution of lymphocyte globulin to the blood proteins.

3. Any stimulus or stress which augments pituitary-adrenal cortical secretion accentuates lymphocyte dissolution and globulin release.

4. It is suggested that one of the major functions of the lymphocytes may be to serve as a medium of distributing throughout the body their important cytoplasmic protein, γ-globulin.

BIBLIOGRAPHY

1. **Moon, H. D.**
   1937. Proc. Soc. Exp. Biol. & Med. **37**: 34.
2. **Selye, H.**
   1937. Endocrinology **21**: 169.
3. **Ingle, D. J.**
   1938. Proc. Soc. Exp. Biol. & Med. **38**: 443.
4. **Ingle, D. J.**
   1940. Proc. Soc. Exp. Biol. & Med. **44**: 174
5. **Reinhardt, W. O., & R. O. Holmes**
   1940. Proc. Soc. Exp. Biol. & Med. **45**: 267.

6. **Selye, H.**
1940. Cyclopedia of Medicine, Surgery & Specialties **15**: 15.
7. **Wells, B. B., & E. C. Kendall**
1940. Proc. Staff Meetings Mayo Clinic **15**: 133.
8. **Dougherty, T. F., & A. White**
1943. Proc. Soc. Exp. Biol. & Med. **53**: 132.
9. **Simpson, M. E., C. H. Li, W. O. Reinhardt, & H. M. Evans**
1943. Proc. Soc. Exp. Biol. & Med. **54**: 153.
10. **Dougherty, T. F., & A. White**
1944. Endocrinology **35**: 1.
11. **White, A., & T. F. Dougherty**
1944. Proc. Soc. Exp. Biol. & Med. **56**: 26.
12. **Dougherty, T. F., & A. White**
1945. Am. J. Anat. **77**: 81.
13. **Sabin, F. R.**
1939. J. Exp. Med. **70**: 67.
14. **Perla, D., & J. Marmorston**
1941. Natural Resistance and Clinical Medicine. Little, Brown & Co. Boston.
15. **McMaster, P. D., & S. S. Hudack**
1935. J. Exp. Med. **61**: 783.
16. **Ehrich, W. E., & T. N. Harris**
1942. J. Exp. Med. **76**: 335.
17. **Grollman, A.**
1936. The Adrenals. Williams & Wilkins Co. Baltimore.
18. **White, A., & T. F. Dougherty**
Unpublished results.
19. **Sayers, G., M. A. Sayers, A. White, & C. N. H. Long**
1943. Proc. Soc. Exp. Biol. & Med. **52**: 200.
20. **Sayers, G., M. A. Sayers, E. G. Fry, A. White, & C. N. H. Long**
1944. Yale J. Biol. & Med. **16**: 361.
21. **White, A., & T. F. Dougherty**
1945. Endocrinology **36**: 16.
22. **Reinhardt, W. O., & C. H. Li**
1945. Science **101**: 360.
23. **Reinhardt, W. O., H. Aron, & C. H. Li**
1944. Proc. Soc. Exp. Biol. & Med. **57**: 19.
24. **Downey, H., & F. Weidenreich**
1912. Arch. f. mikr. Anat. &. Entwgesch. **80** (I): 306.
25. **White, A., & T. F. Dougherty**
1945. Endocrinology **36**: 207.
26. **Dougherty, T. F., J. H. Chase, & A. White**
1944. Proc. Soc. Exp. Biol. & Med. **57**: 295.
27. **Dougherty, T. F., A. White, & J. H. Chase**
1945. Proc. Soc. Exp. Biol. & Med. **59**: 172.
28. **Dougherty, T. F., J. H. Chase, & A. White**
1945. Proc. Soc. Exp. Biol. & Med. **58**: 135.
29. **White, A., & T. F. Dougherty**
1945. Fed. Proc. Am. Soc. Biol. Chemists **4**: 109.
30. **Dougherty, T. F., A. White, & J. H. Chase**
1944. Proc. Soc. Exp. Biol. & Med. **56**: 28.

30a. **Chase, J. H., A. White, & T. F. Dougherty**
1946. J. Immunol. **52**: 101.
31. **Cuthbertson, D. P.**
1942. Lancet **I**: 433.
32. **Dougherty, T. F., & A. White**
1945. Anat. Rec. **91**: 7. Suppl.
33. **Whipple, G. H.**
1942. Am. J. Med. Sci. **203**: 477.
34. **Whipple, G. H., & S. C. Madden**
1944. Med. **23**: 215.
35. **Cannon, P. R., E. M. Humphreys, R. W. Wissler, & L. E. Frazier**
1944. J. Clin. Invest. **23**: 601.
36. **Cannon, P. R.**
1942. J. Lab. & Clin. Med. **28**: 127.
37. **Ehrich, W. E., & T. N. Harris**
1945. Science **101**: 28.

DISCUSSION OF THE PAPER

Dr. Heidelberger:

Again, I feel that, while Dr. White's results supply evidence for an additional and, perhaps, major source of antibodies, they do not exclude the phagocytic cells of the reticuloendothelial system. Lymphocytes are not the only cells that throw off surface films, and Dr. Florence Sabin[1] has shown very clearly that macrophages containing antigen throw off such films at about the time of appearance of antibodies in the serum.

Dr. McMaster:

In the work of Dr. Sabin, those Kupffer cells which had ingested colored antigen were later observed shedding cytoplasm into the circulation, as if engaged in the liberation of antibody. Dr. Sabin has suggested lately that the Kupffer cells, instead of forming antibodies, might have been engaged in the preparation of antigen for its take-up by lymphocytes.

Dr. Dougherty:

There would seem to be some question as to whether the lymphocytes undergoing dissolution yield their cytoplasmic material to the blood directly, or whether these cells are first phagocytized and their cytoplasmic constituents subsequently discharged by macrophages.

Cytological and physiological evidence indicates that lymphocytes yield portions of their cytoplasm directly to the lymph. The shedding or budding of lymphocyte cytoplasm and the disintegration of the nucleus resulting in dissolution of lymphocytes occurs shortly after hormone treatment. It is observed in the germinal centers, within an hour after injection of adrenal cortical extracts or certain adrenal cortical steroids. The ingestion of the free nuclear matter does not take place until some time later (3–6 hours). Although some macrophages are observed in the lymphoid tissues earlier than this, they are not actively phagocytising nuclear particles or whole lymphocytes. Most of the dissolution of lymphocytes seems to be over by the time the macrophages actively begin to ingest the scattered nuclear material. However, there are several different types of histiocytes, which become evident in lymphoid structures following adrenal cortical stimulation. The cytoplasm of fixed reticulum cells contains a peculiar dark blue granular substance which could very possibly be ingested lymphocyte cytoplasm. It is interesting that this material is not seen in the cytoplasm of other phagocytic cells.

[1] J. Exp. Med. **70**: 67. 1939.

Physiological evidence concerning the direct release of lymphocyte cytoplasm is yielded by the fact that the increase in γ-globulin and antibodies in the blood occurs at the time lymphocytes are undergoing dissolution, and before there is much phagocytosis of disintegrating or whole lymphocytes.

Dr. Merrill Chase (*Rockefeller Institute for Medical Research, New York, N. Y.*):

All of us would probably agree that, with regard to the ingestion of soluble antigens by lymphocytes and the direct elaboration of antibodies by these cells, there would be no special conceptual difficulty. In the case of the formation of antibodies to insoluble antigens (*e.g.*, coagulated ovalbumin), it may be well to consider with more caution the idea that phagocytic cells—Kupffer cells, clasmatocytes, and the like—may act on the antigen, primarily, and, in turn, pass over the proper soluble intermediates to lymphocytes, prior to actual antibody elaboration.

Cleavage by proteolytic enzymes, with reduction in particle size, is destructive of the antigenicity both of native proteins and denatured (coagulated) proteins (although, with special serological systems, species specificity is still found in the proteoses[1]). It may be in place to recall that Rothen and Landsteiner studied the serological reactivity of heat-denatured but soluble ovalbumins of a series of birds, against antibody developed to firmly heat-coagulated hen ovalbumin.[2] Denaturation of the hen ovalbumin altered the initial specificity of the native protein, so that the antibodies which were formed scarcely reacted at all with native ovalbumin; yet, these antibodies reacted with the soluble heat-denatured ovalbumins of the several species, giving cross-reactions of the same relative intensities as occur between the respective native ovalbumins. The species specificity had not, then, been abolished by denaturation, but had been established, as it were, on a new antigenic level. But antibodies of this sort would not have been produced, had proteolytic enzymes acted on the coagulated hen ovalbumin for only a short time. Therefore, to attribute the production of antibody to lymphocytes, in this case, one must conceive that the cells capable of direct ingestion of insoluble materials can reduce the coagulum to the size of "soluble particles" and release these, without proteolytic alteration, before the lymphocytes are to receive and use them as antigenic matrix.

We may also ask whether it is only lymphocytes that produce normal globulin. There is one observation of preliminary character, which, although needing confirmation and study, may be mentioned here. Landsteiner and Parker[3] had found that chicken connective tissue cells (fibroblasts), maintained on rabbit plasma medium, continued, over a period of nearly 8 months (35 passages), to elaborate substances which, by their serological reactivity, seemed to be chicken serum proteins. I happened to be present when further and unpublished experiments were made. A preliminary result, secured by an indirect testing method, made it appear probable that both serum globulin and serum albumin were being elaborated by these cultures of fibroblasts. Of course, γ-globulin might not be represented here.

These observations are not made with the idea of detracting from pursuit of the newer concept of the function of lymphocytes, but seem to merit attention in the development of a theory.

---

[1] **Landsteiner, K., & M. W. Chase.** Proc. Soc. Exp. Biol. & Med. **30:** 1413–1415. 1933.

[2] **Rothen, A., & K. Landsteiner.** Serological reactions of protein films and denatured proteins. J. Exp. Med. **76:** 437–450. 1942. **Landsteiner, K., & J. van der Scheer.** Ibid. **71:** 445–454. 1940.

[3] **Landsteiner, K., & R. C. Parker.** Serological tests for homologous serum proteins in tissue cultures maintained on a foreign medium. J. Exp. Med. **71:** 231–236. 1940.

PLATE 14

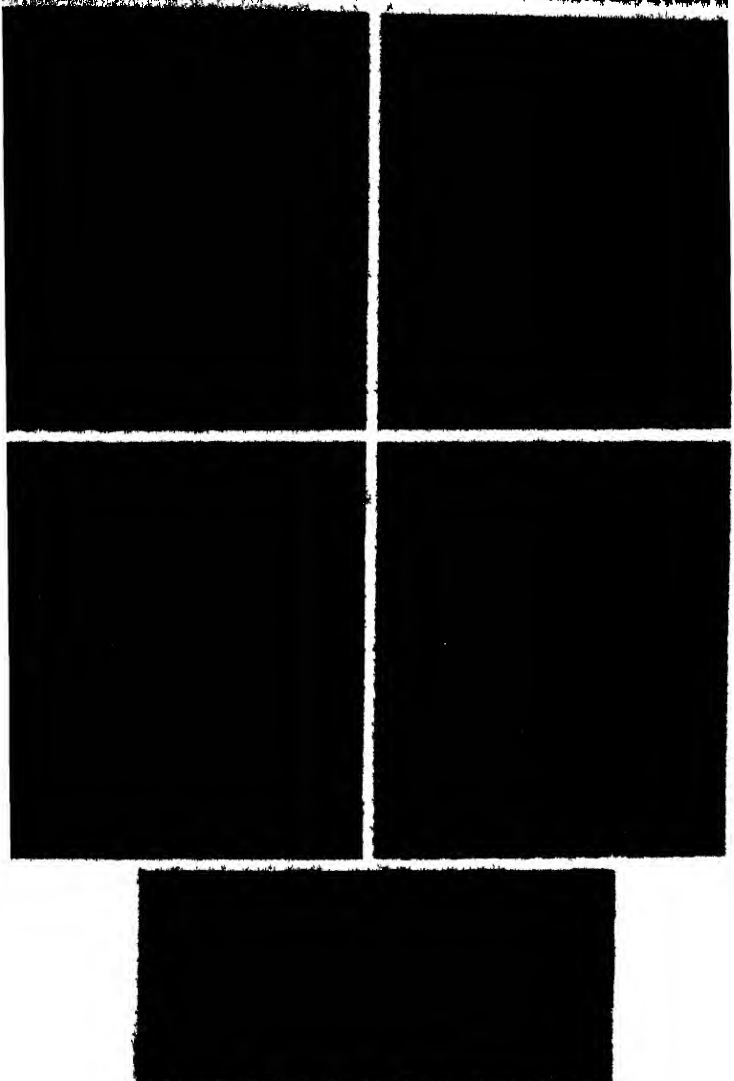
FIGURE 1. Inguinal lymph node of mouse, 3 hours following subcutaneous injection of adrenotrophic hormone. The node contains so much fluid that the pycnotic lymphocytes are pushed into clusters. Fixed in Bouin, sectioned at 5 micra, and stained with haematoxylin and eosin. (x 90.)

FIGURE 2. Thymus of mouse, 3 hours after adrenotrophic hormone injection. The thymus is edematous, and many small clusters of nuclear debris are scattered throughout the tissue. Free nuclear debris is usually not phagocytized until 4 to 6 hours after hormone injection. Tissues treated in same manner as in Fig. 1. (x 90.)

FIGURE 3. Numerous macrophages, having large amounts of clear cytoplasm and containing phagocytized nuclear debris. The clear cytoplasm gives the thymus a "pitted" appearance. The cells about the macrophages are degenerating lymphocytes. This is a rabbit thymus, 9 hours following injection of adrenotrophic hormone. Fixed in Zenker-formol, sectioned at 5 micra, and stained with haematoxylin and eosin. (x 475.)

FIGURE 4. Edematous germinal center of Peyer's patch of mouse, 9 hours following the injection of adrenotrophic hormone. Note the large amount of free nuclear debris. Many of the lymphocytes in the collar in the right hand side of the figure are markedly shrunken. Fixed in Bouin, sectioned at 5 micra, and stained with haemotoxylin and eosin. (x 350.)

FIGURE 5. Lymphocytes in dry film preparation of lymph taken from edematous mesenteric lymph node of rabbit, 6 hours after injection of adrenal cortical steroids in oil. Shrunken lymphocytes devoid of cytoplasm, and lymphocytes having cytoplasmic buds are numerous in this preparation. The large lymphocyte is an immature form which generally does not undergo dissolution under the influence of the adrenal cortical steroids. However, the cytoplasm of such cells becomes heavily basophilic. Air-dried smear stained with May-Grünwald Giemsa. (x 1550.)



...: LYMPHOCYTES AND GLOBULIN

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES

VOLUME XLVI, ART. 9. PAGES 883–992

NOVEMBER 8, 1946

BLOOD GROUPING*

BY

WILLIAM C. BOYD, J. W. CAMERON, L. K. DIAMOND, PHILIP LEVINE, M. MELIN, J. L. ONCLEY, LOUIS PILLEMER, D. A. RICHERT, EVE B. SONN, A. S. WIENER, AND ERNEST WITEBSKY

CONTENTS

* This series of papers is the result of a Conference on Blood Grouping held by the Sections of Biology, and Physics and Chemistry, of The New York Academy of Sciences, May 18 and 19, 1945

Publication made possible through a grant from the Conference Publications Revolving Fund.

INTRODUCTION TO THE CONFERENCE ON BLOOD GROUPING

By William C. Boyd

Boston University School of Medicine, Boston, Massachusetts

There is no necessity to dilate on the importance which the subject of blood grouping has acquired. The events of the Second World War alone have been sufficient to underline it. The life-saving value of plasma and plasma derivatives, and of whole blood, has been sufficiently dramatic to bring the importance of the subject home to scientist and layman alike. Our military authorities have finally become interested in blood grouping.

At the present conference, we are privileged to have papers from some of the outstanding contributors to the study of problems of blood groups; in particular, some who have worked on the problem of war-time transfusions; some who have worked on the production of blood grouping reagents for the armed forces; and some who are investigating the newer problems which have arisen in connection with the newly discovered Rh factor. We have most of the men who did the leading work on these lines with us at this conference.

Like the rest of you, I deeply regret the absence of Dr. Karl Landsteiner, the pioneer of pioneers in this field, whose brilliant and fruitful career was cut short by death. If some, or even one, of us here can achieve and maintain the impartial search for truth, complete devotion to science, and unremitting industry which were such conspicuous features of Dr. Landsteiner's character throughout his long and productive scientific career, we may be assured that our subject will not decline in importance.

ISOLATION AND PURIFICATION OF BLOOD GROUP A AND B SUBSTANCES; THEIR USE IN CONDITIONING UNIVERSAL DONOR BLOOD, IN NEUTRALIZING ANTI-Rh SERA, AND IN THE PRODUCTION OF POTENT GROUPING SERA

By Ernest Witebsky

Departments of Bacteriology and Immunology, University of Buffalo Medical School, and Buffalo General Hospital, Buffalo, N. Y.

The chemical nature of the blood group specific substances has been the subject of discussion for many years. Proteins, carbohydrates, and lipoids have been considered to be carriers of blood group specific properties. We are dealing with a rather perplexing situation. Undoubtedly, alcohol soluble substances of lipoid character play an important role in the group specific stigmatization of red blood cells, as well as of the tissues. In contrast, blood group specific substances occurring in high concentrations in certain secretions of the human body, such as saliva, gastric juice, and amniotic fluid, are carbohydrate-like in nature. Carbohydrate fractions exhibiting the A, B, and O properties, however, have not, as yet, been isolated from the blood cells themselves, in spite of the considerable effort that has been spent on this problem. Some peptone preparations, as well as pepsin, are known to contain various amounts of the A factor. The B substance is completely absent in these commercial preparations. Highly potent A substance can be isolated from the gastric mucosa of hogs, a good percentage of which contains the A factor, but no trace of the B factor. Most investigators, therefore, limited their investigations on the chemical nature of the blood group specific substance to the A property. In 1940, using a method similar to Goebel's procedure, namely, multiple alcoholic precipitation, a carbohydrate preparation exhibiting B property was isolated from the gastric juice of human beings, by Klendshoj and the author.

From the standpoint of mass production, it seemed advisable to look for sources of blood group specific substances outside of the human body. There is no difficulty, as far as the A substance is concerned,

but there is no easily available source of B substances. Fractions of the B factor are frequently found in smaller laboratory animals. Among the larger ones, the horse was found to contain a B substance similar to the human one. In the beginning of our investigations, the saliva of horses was used as a source of the B substance, but this did not prove to be a reliable one. Further studies regarding the distribution of the B factor within the body of the horse revealed the gastric mucosa to be rich in B substance. As a matter of fact, there was a definite quantitative difference in the content of the B substance of the saliva of horses, on one hand, and the gastric mucosa, on the other. This difference sometimes assumed qualitative proportions, inasmuch as, occasionally, carbohydrate fractions exhibiting B properties could be isolated from the gastric mucosa of the horse, which could not be demonstrated in the saliva of the horse during its lifetime. Most of the mucous membranes of the horse's stomach contained varying amounts of both A and B substances. Only in rare instances was pure B found. The mucous membranes of several horses have to be pooled in order to obtain sufficiently large amounts of the final preparation. For that reason, the product obtained from horses is not a B substance, but an AB substance. The carbohydrate fraction exhibiting blood group specific properties is obtained from the horse stomach mucosa by digestion with pepsin. The carbohydrate fraction is then isolated by multiple alcohol precipitation. The final product, a whitish powder, is easily dissolved in water. At the present time, some pharmaceutical houses are preparing the blood group specific substances on a larger scale. Two different substances are available: (1) The A specific substance isolated from the hog stomach; (2) the AB specific substance isolated from the horse stomach. These preparations are chemically free of protein. Their failure to sensitize guinea pigs, as well as other properties, is subject to rigid specifications recently set up by a committee of the National Research Council.

The following experiment shows the efficacy of the isolated blood group specific substances in suppressing isoagglutination. Decreasing amounts of

(a) A specific substance (1% solution), and

(b) AB specific substance (1% solution) (volume 0.1 cc.)

were mixed with 0.1 cc. of a 1:20 diluted serum of group A. After incubating the mixtures for 15 minutes at room temperature, 0.1 cc. of human cells belonging to group B (2% suspension) was added. The

tubes were thoroughly shaken and allowed to stand at room temperature for an additional 15 minutes. Then the tubes were centrifuged, for 2 minutes, at medium speed The resulting agglutination can be seen in TABLE 1.

TABLE 1

INHIBITION OF AGGLUTINATION OF B CELLS BY SERUM OF GROUP A PREVIOUSLY TREATED WITH BLOOD GROUP SPECIFIC SUBSTANCES

Substance to be tested (1%)		A substance	AB substance
(1) Undiluted	0.1 cc	++++	−
(2) 1:3	0.1 cc	++++	−
(3) 1:9	0.1 cc	++++	−
(4) 1:27	0.1 cc	++++	−
(5) 1:81	0.1 cc	++++	−
(6) 1:243	0.1 cc	++++	−
(7) 1:729	0.1 cc	++++	−
(8) 1:2187	0.1 cc	++++	±
(9) 1:6561	0.1 cc	++++	+
(10) 1:19683	0.1 cc	++++	++
(11) 1:59049	0.1 cc	++++	+++
(12) 0	0.1 cc	++++	++++

− = no agglutination,
± = faint agglutination,
+ = slight agglutination,
++ = marked agglutination,
+++ = strong agglutination,
++++ = very strong agglutination.

The agglutination of B cells by a serum of group A is definitely inhibited by the AB specific substance, but not at all by the A specific substance. Dilutions of 1:729 (tube 7) of the 1% solution of the AB substance still completely inhibit agglutination of B cells by the serum of group A, while partial inhibition of agglutination is still traceable down to the 11th tube containing a dilution of 1:59,049 of the 1% stock solution. This is true, in spite of the fact that the control tube shows a strong ++++ agglutination, indicating that all cells have been agglutinated in one clump which cannot be shaken out. If the serum used in TABLE 1 had been diluted a hundred times, instead of twenty times, the extent of inhibition of agglutination by the blood group specific substances would have been considerably higher.

From the practical standpoint, commercially available A and B specific substances have been applied, at least, for the following three entirely different purposes:

1. NEUTRALIZATION OF ISOANTIBODIES ANTI-A AND ANTI-B IN O BLOOD

The attitude of the medical profession towards the conception of the universal donor has changed several times, during the past years. Some years ago, it was felt by many authors that, under certain circumstances, the isoagglutinins anti-A or anti-B, when present in high titer, would constitute a source of hemolytic transfusion reaction, in some instances. The term, "*dangerous universal donor,*" was introduced for high-titered O blood. With the discovery of the Rh factor, hemolytic transfusion reactions previously attributed to the dangerous universal donor were now explained on the Rh incompatibility of the transfused blood. There was a tendency to neglect completely the significance of the transfusion of incompatible isoagglutinins anti-A or anti-B in individuals corresponding to the respective blood groups. However, ample evidence is now available to show that high-titered universal blood can cause hemolytic and even fatal transfusion reactions in patients belonging to blood groups other than blood group O, especially in A and AB patients. Considering the fact that the red blood cells of O blood are not agglutinated, nor dissolved, by the isoantibodies anti-A and anti-B, the only objection against the indiscriminate use of O blood consisted, therefore, in the existence of high titered isoagglutinins anti-A and anti-B in the O-donor's plasma. In order to reduce the titer of isoantibodies in O blood, a few milligrams of the blood group specific substances are dissolved in saline solution and added to one pint of O blood, either as soon as the blood is collected or before the transfusion is given. The latter procedure is feasible, because the interaction of the isoagglutinins and their corresponding antigens is practically an immediate one. The addition of the isolated blood group specific substances reduces the isoantibody titer in O blood considerably, eliminating 75 to 95% of the antibodies present. It does not seem necessary to reduce the antibody titer to zero.

An example, illustrating what happens if one vial containing a few milligrams of blood group specific substances dissolved in 10 cc. of saline solution, or a fraction of such a vial, is added to one pint (500 cc.) of O blood, is given in TABLE 12. Half a vial (5 cc.) of the preparation J72-10 (Lilly) was added to one pint of O blood. A small specimen of plasma was taken with a syringe, under sterile conditions, before the substances were added; and a second specimen, 15 minutes after the substances were added to the blood. The experiment itself was carried out in the following way: Decreasing dilutions of plasma

(volume 0.2 cc.) were mixed with 0.2 cc. of a 2% suspension of washed A cells, and 0.2 cc. of a 2% suspension of washed B cells, respectively. The tubes were shaken thoroughly, and kept at room temperature for 15 minutes. They were then centrifuged at medium speed and read for agglutination.

TABLE 2

AGGLUTINATION OF A CELLS AND B CELLS BY HUMAN PLASMA BELONGING TO BLOOD GROUP O

	I Before the addition of blood group specific substances		II After the addition of blood group specific substances (One-half vial)	
Plasma group O	A cells	B cells	A cells	B cells
(1) Undiluted	++++	++++	+	+
(2) 1:2	++++	++++	±	+
(3) 1:4	++++	++++	−	±
(4) 1:8	++++	++	−	−
(5) 1:16	++	−	−	−
(6) 1:32	+	−	-	−
(7) 1:64	±	−	−	−
(8) 0	−	−	−	−

As TABLE 2 shows, the addition of as little as half a vial of blood group specific substances (Lilly) has reduced the anti-A and anti-B titer of the plasma to a considerable extent. The potency of the blood group specific substances is checked in that way. The A content of the vials is higher than their B content, mainly because, frequently, the isoagglutinins anti-A in O blood are stronger than the anti-B.

Another example showing the potency of the substances is given in TABLE 3, in which plasma of group O with an unusually high anti-B antibody was examined. The experiment was carried out in the following way: Decreasing dilutions of plasma group O, volume 0.2 cc., were mixed with 0.2 cc. of a 2% suspension of washed cells belonging to blood group A. A small specimen of plasma was obtained before the addition of the blood group specific substances, as well as half an hour after the addition of one-third of a vial of preparation J 72–22 (Lilly). Otherwise, the experiment was carried out in exactly the same way as that given in TABLE 2.

The addition of one-third of a vial of blood group specific substances was sufficient to reduce the titer of this plasma from 2048 to 128. It should be stressed, however, that the addition of blood group specific

TABLE 3
AGGLUTINATION OF A CELLS BY HIGH TITERED HUMAN PLASMA BELONGING TO BLOOD GROUP O

	I Before the addition of blood group specific substances	II After the addition of blood group specific substances (One-third vial)
Plasma dilutions	A cells	A cells
(1) 1:2	++++	++++
(2) 1:4	++++	+++
(3) 1:8	++++	+++
(4) 1:16	++++	++
(5) 1:32	++++	+
(6) 1:64	++++	±
(7) 1:128	++++	±
(8) 1:256	+++	−
(9) 1:512	++	−
(10) 1:1024	+	−
(11) 1:2048	±	−
(12) 0	−	−

substances results not only in the reduction of the end titer, but also in the reduction of the strength and quality of agglutination in the remaining dilutions which still show some agglutination. Isohemolysis, also, is completely inhibited by the addition of the blood group specific substances.

2. COMPLETE SUPPRESSION OF ISOAGGLUTININS ANTI-A AND ANTI-B IN ANTI-Rh SERUM

Another practical application of the purified blood group specific substances consists in the elimination of isoagglutinins anti-A and anti-B in anti-Rh sera used for the determination of the presence of the Rh factor. In this case, the isoagglutinins anti-A and anti-B have to be completely neutralized, in order to allow the anti-Rh antibody to act specifically against the Rh positive cells. Therefore, the isoagglutinins anti-A and anti-B have to be reduced to zero. Consequently, a surplus of A and B substances should be added, in order to be sure that no trace of the isoagglutinins anti-A and anti-B has been left, which would result in mistaken determination of the Rh factor. A typical example of such a procedure follows: Serum (Car) belonging to blood group A was received from a woman who had given birth to an erythroblastotic child. This serum was mixed in the following proportions with AB substances (vial Lilly):

I. Serum (Car) without the addition of the AB substances (control).
II. 16 parts of serum (Car) plus 1 part of AB substance.
III. 8 parts of serum (Car) plus 1 part of AB substance.
IV. 4 parts of serum (Car) plus 1 part of AB substance.
V. 2 parts of serum (Car) plus 1 part of AB substance.
VI. 1 part of serum (Car) plus 1 part of AB substance.

The mixtures were kept for 15 minutes in the icebox. Then the following experiment was carried out: Decreasing dilutions of the above mixtures (I–VI) (volume 0.1 cc.) were mixed with: (a) 0.1 cc. of a 2% suspension of Rh positive group B cells; (b) 0.1 cc. of a 2% suspension of Rh negative group B cells. The tubes were kept for one hour at room temperature, as this serum acted at room temperature almost as well as at 37° C. Then the tubes were centrifuged at medium speed and read for agglutination.

TABLE 4 shows that serum (Car) agglutinates human cells belonging to blood group B almost to the same extent, irrespective of whether they are Rh positive or Rh negative. The addition of small amounts of blood group specific substances reduces the isoagglutinin anti-B which is present in the patient's serum. The degree of this neutralization depends upon the amount of the blood group specific substances used. The addition of one part of AB substances to 16 and 8 parts of serum (Car), respectively, leads to an incomplete elimination of the isoagglutinin anti-B. One part of AB substances added to 4 parts of the serum completely neutralizes it. For practical purposes, it would seem wise, however, to use an even larger amount of blood group specific substances, possibly equal amounts of serum and blood group specific substances, in order to be sure that all traces of the agglutinin anti-B have disappeared. After the elimination of the isoagglutinin anti-B, the anti-Rh antibody manifests itself, and the experiment further shows that the blood group specific substances do not interfere with the activity of the Rh antibody.*

3. PRODUCTION OF POTENT GROUP TEST SERUM

The purified A and B specific substances are useful in the production of potent blood grouping test serum. In the past, an increase in the

* However, it should be mentioned that the vials containing A and B substances as prepared by commercial houses contain traces of phenol or similar antiseptics which might, on long contact with anti-Rh serum, adversely affect the anti-Rh antibody. It is, therefore, recommended either to neutralize anti-Rh serum shortly before it is used, or use group specific substances which do not contain antiseptics which might destroy the Rh antibody.

TABLE 4

AGGLUTINATION OF RH POSITIVE AND RH NEGATIVE B CELLS BY AN ANTI-RH GROUP A SERUM TREATED WITH VARIOUS AMOUNTS OF BLOOD GROUP SPECIFIC SUBSTANCES

Serum (Car) Group A		I		II		III		IV		V		VI	
		a Rh+ B	b Rh− B	a Rh+ B	b Rh− B	a Rh+ B	b Rh− B	a Rh+ B	b Rh− B	a Rh+ B	b Rh− B	a Rh+ B	b Rh− B
(1) Undiluted	0.1 cc	++++	++++	+++	++	++++	+	+++	−	+++	−	+++	−
(2) 1:2	0.1 cc	++++	++++	++++	++	+++	±	+++	−	+++	−	+++	−
(3) 1:4	0.1 cc	++++	++++	+++	+	+++	−	+++	−	+++	−	++	−
(4) 1:8	0.1 cc	+++	+++	+++	±	+++	−	++	−	++	−	++	−
(5) 1:16	0.1 cc	+++	++	++	−	++	−	++	−	+	−	+	−
(6) 1:32	0.1 cc	++	+	++	−	+	−	+	−	+	−	±	−
(7) 1:64	0.1 cc	+	±	+	−	±	−	−	−	−	−	−	−
(8) 1:128	0.1 cc	±	−	−	−	−	−	−	−	−	−	−	−
(9) 0	0.1 cc	−	−	−	−	−	−	−	−	−	−	−	−

I: Native Serum (Car) not neutralized
II: 16 parts Serum (Car) and 1 part AB substances (Lilly)
III: 8 parts Serum (Car) and 1 part AB substances (Lilly).
IV: 4 parts Serum (Car) and 1 part AB substances (Lilly).
V: 2 parts Serum (Car) and 1 part AB substances (Lilly)
VI: 1 part Serum (Car) and 1 part AB substances (Lilly)

isoagglutinin titer, following the erroneous transfusion of incompatible blood, has been observed in several instances. Recently, the injection of relatively large amounts of pooled plasma has been reported, by Aubert, Boorman, and Dodd, to result in the increase of the isoagglutinin titer. The presence of small amounts of soluble A and B substances in pooled plasma was considered to be the cause of the increase.

The amount of test serum being used, at the present time, in laboratories connected with transfusion services is very large. Lack of sufficiently large amounts of potent test serum during the war presented a major problem. Obviously, the better the test serum, as far as the degree and the rapidity of the agglutination are concerned, the fewer mistakes occur in routine blood group determinations.

The intravenous injection of the isolated group specific substances leads to an increase in the agglutinin titer, which is sometimes rather remarkable. At the beginning of these investigations, as much as the content of one vial (10 cc.) was given to each of several volunteers. Their isoagglutinin titer dropped immediately following injection, and did not return to normal for two to five days, depending upon the amount of material given. However, after a week or so, the donor's agglutinin titer rose above the original strength and became very high. FIGURE 1 illustrates the isoagglutinin titer of an individual belonging to group A, treated in the described way and tested at different intervals for a period of several weeks.

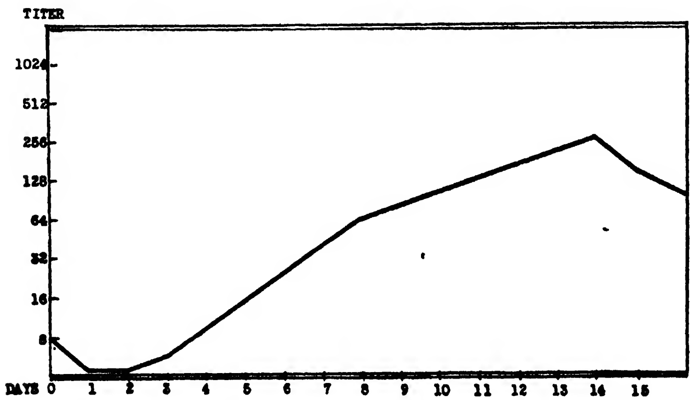


FIGURE 1. Patient M U, blood group A, treated with Eli Lilly's A & B substance Curve of anti-B agglutinins.

After about 12 to 14 days following the injection of the blood group

specific substances, the isoagglutinin anti-B has been increased about thirty times (FIGURE 1). Individuals whom we rejected were examined over a period of one to two years. Their agglutinin titer remained elevated at the end of the observation period, as compared with the original titer, though these agglutinins showed a tendency to decrease in potency. Further investigations showed intravenous injections of minute amounts, as low as 0.1 to 0.2 cc. of the solution of blood group specific substances as prepared commercially, containing as little as one-tenth of a milligram, would result in a 10- to 100-fold increase in the original isoagglutinin titer. Of course, the injection of small amounts such as 0.1 to 0.2 cc. of blood group specific substances did not reduce at all, or only slightly, the original isoagglutinin titer.

Test sera produced by immunization of volunteers with the blood group specific substances not only show a high end titer, but the character and quality of the agglutination, also, are changed. The majority of donors treated with A and B substances reacted with the described increase in isoagglutinin titer. A minority failed to respond at all, or showed only a negligible increase. Potent test sera obtained by treatment of donors with A and B substances can frequently be diluted considerably, before being used, and might still be more potent, after proper dilution, than most of the normal sera used as test sera. The time of agglutination is frequently shortened. Cells of the A_2 and A_2B groups are strongly agglutinated by such an immune test serum. Even A_3 cells which, sometimes, are not agglutinated, or only to a very weak extent, by the average normal test serum, are strongly agglutinated by the serum of individuals treated with blood group specific substances. It should be pointed out that, because of the extraordinary antigenic stimulation exerted by minute amounts of purified type A and B substances, one vial might be sufficient to immunize a good number of A, as well as B, individuals, leading to a large supply of potent test serum of group A and group B. Persons belonging to blood group A, when injected with AB substances, respond by an increase in the antibody anti-B, the A substance being without any effect on A individuals. B individuals injected with the same material react with an increase in the anti-A titer, as the B substance present in the mixture used for treatment is without antigenic effect in people belonging to group B.

Investigations on the immunochemical nature of the blood group specific substances, originally undertaken only for theoretical purposes, have led to several possible applications in the field of blood transfusion.

DISCUSSION OF THE PAPER

Dr. S. H. Polayes (*Brooklyn, N. Y.*):

Dr. Witebsky states that he is convinced of the specificity of the O properties. I would like to ask him if he has been able to identify a specific anti-O agglutinin which might conceivably have resulted from transfusion of group O blood into a recipient of one of the other groups? Despite the belief that O antigen is present in the blood of all groups, making it difficult to immunize an O with O blood, may there not be a variant of the O antigen, analogous to the subgroup of A, for example, which variant may be sufficiently specific to induce its own specific agglutinin and, thus, immunize the recipient, so that, on subsequent transfusion, a reaction might occur? This would explain reactions to the so-called universal donors which are not based on Rh incompatibility. It might even explain the reaction in the case just cited by Dr. Philip Levine, if, in his case, the anti-A and anti-B agglutinins of the O donor's blood could be exonerated, as, indeed, the advocates of the indiscriminate use of the universal donor would have us do. In Dr. Levine's case, a considerable time did elapse between the first and final transfusions, a period sufficient to permit the development of an immune body to the donor's blood. Although the natural supposition would be that the anti-A and anti-B agglutinins were responsible for the reaction, the other is still a possibility to be considered, for we also get unexpected reactions, even when the recipient is a group O individual.

Dr. William C. Boyd (*Boston University, School of Medicine, Boston, Mass.*):

In reply to the remarks of Dr. Polayes, it seems to me more likely that the reactions observed to transfusions of group O, "universal donor," blood are due to the action of the anti-A or anti-B agglutinins in the transfused blood on the cells of the recipient.

Dr. Elvin A. Kabat and Miss Ada E. Bezer (*Columbia University, New York, N. Y.*):

The Estimation of Isoantibodies by the Quantitative Precipitin Reaction

Precipitin reactions have not been known to occur between purified blood group substances and their homologous isoantibodies present in human serum. The demonstration by Dr. Witebsky and his collaborators,[1] that the purified A and B specific substances are antigenic in man, has recently made available sera containing high titers of isoagglutinins. By applying the micro-quantitative precipitin methods developed by Heidelberger and MacPherson,[2] and using a preparation of the blood group A, prepared by the phenol method as described by Morgan and King,[3] it has been possible to obtain precipitin reactions with human sera containing anti-A, and to estimate the amounts of antibody in these sera on a weight basis. Using sufficiently large volumes of serum, amounts of precipitate suitable for analysis could be obtained with sera of normal individuals of groups O and B, whereas negligible amounts of precipitate were in group A and AB sera.

On immunization of group O individuals, with the A and B substances manufactured and supplied by Eli Lilly and Co., increases in titer and in capacity of the serum to neutralize A substance were accompanied by increases in the amounts of precipitin. A typical instance is given below (Table 1):

Table 1

Blood group O	Titer	Combining capacity for A substances per 0.1 ml. serum	Precipitable anti-A nitrogen
		μg	μg/ml.
Before immunization	4–8	0.5	3.2
After immunization	512	90	54

Individuals of blood group A formed small amounts of precipitin (1–5 μgN/ml.) with A substance, after immunization with A and B substances. These antibodies, however, could be shown not to be related to the A substance, since all of the A substance remained in the supernatant, after removal of the precipitate.

The precipitin reaction obtained with A substance and the sera of immunized individuals of blood group O has been shown to be similar, in its quantitative course, to that of other antigen-antibody systems. If increasing amounts of A substance are added to a given volume of serum, larger quantities of antibody nitrogen are precipitated, until a maximum is reached. Tests on supernatants show that the agglutinin titer of the supernatant decreases with the precipitation of increasing amounts of antibody nitrogen, that maximum precipitation of antibody occurs at the point where A substance first appears in excess in the supernatant, and that anti-A and A substance do not coexist in the same supernatant solution. Data are given in TABLE 2.

TABLE 2

ADDITION OF INCREASING AMOUNTS OF A SUBSTANCE TO 1.0 ML. OF SERUM OF GROUP O FROM AN INDIVIDUAL PREVIOUSLY IMMUNIZED WITH A AND B SUBSTANCES

A substance added	Antibody N precipitated	Tests on Supernatants: Anti-A titer	Tests on Supernatants: Amount A substance in supernatant*
μg	μg		μg
Total volume, 3 ml.; original titer of serum, 512			
25	18.1	32	0
50	37.2	8	0
75	43.6	2	0
100	47.6	1	0
150	55.8	0	1–2
200	58.5	0	15

* Assayed by measuring capacity to inhibit agglutination.

The ability to obtain precipitin reactions between blood group A substance and its homologous antibody should prove of considerable value, in following the course of the purification of the substance and in providing a precise method for standardizing isoagglutinins and studying variations in isoagglutinin levels.

REFERENCES

1. **Witebsky, E., N. C. Klendshoj, & C. McNeil**
   1944. Proc. Soc. Exp. Biol. & Med. 55: 165.
2. **Heidelberger, M., & C. F. C. MacPherson**
   1943. Science 97: 405; 98: 63.
3. **Morgan, W. T. J., & H. R. King**
   1943. Biochem. J. 37: 640.

METHODS FOR THE PREPARATION OF ANTI-A, ANTI-B, AND ANTI-Rh ISOAGGLUTININ REAGENTS*†

By J. L. Oncley, M. Melin, J. W. Cameron, D. A. Richert and L. K. Diamond

Department of Physical Chemistry, Harvard Medical School, and Blood-Grouping Laboratory, Children's Hospital, Boston, Massachusetts

For several years, we have been engaged in the study of the fractionation of human plasma, and in the development of large-scale methods for the production of certain of the plasma proteins. As a result of these investigations, it is now possible to separate the proteins into a considerable number of fractions, many of which have been demonstrated to contain certain components of clinical usefulness.[1-3] In addition to the isoagglutinins, these products of plasma include Normal Serum Albumin (Human), Serum γ-Globulin (Immune Serum Globulin, Human), Fibrin Foam and Thrombin (Human), Fibrinogen, and Fibrin Film.

In the Spring of 1942, the anti-A and anti-B isoagglutinins were identified in certain plasma fractions by Dr. W. C. Boyd, who also studied various active sub-fractions and suggested their possible usefulness as reagents for blood grouping. Later that year, investigations under Colonel G. R. Callender, M.C., U.S.A., were begun at the Army Medical School, Washington, D.C., resulting in a method of preparing the anti-A reagent by fractionation of human plasma from random donors of group B, and the anti-B reagent similarly from group A plasma.[4] Subsequently, Colonel Callender requested that the research on isoagglutinin production be carried out at the Department of Physical Chemistry of the Harvard Medical School. Lieutenants L. Pillemer and J. Elliott, and Sergeant M. C. Hutchinson, from the Army Medical School, were sent to Boston to integrate their studies with the other work on human plasma fractionation. A report on the properties of materials so prepared was published by Pillemer, Oncley, Melin, Elliott, and Hutchinson,[5] and a study of the appraisal of these products was made by a group of consultants, col-

* This work has been carried out under contracts recommended by the Committee on Medical Research between the Office of Scientific Research and Development, and Harvard University.
† This paper is Number 55 in the series, "Studies on Plasma Proteins," from the Harvard Medical School, Boston, Massachusetts, on products developed by the Department of Physical Chemistry from blood collected by the American Red Cross.

laborating with the Subcommittee on Blood Substitutes of the National Research Council.[6] As a result of these studies and of later investigations directed toward improving the yields and the qualities of the reagents, satisfactory preparations of the anti-A and the anti-B isoagglutinins from human plasma can now be made available at low cost, as products of a plasma fractionation procedure.

Early studies indicated that the concentration of the anti-Rh agglutinins involved special problems not encountered in the case of the other isoagglutinins, and it was necessary to find procedures which would yield satisfactory reagents for blood grouping work with these anti-Rh antibodies. This has been accomplished, and methods of increasing the potency and stability of anti-Rh reagents have been found.

A. CHOICE OF PLASMA

Anti-A and Anti-B Isoagglutinins

The anti-A or anti-B isoagglutinin content of the plasmas of different human subjects of the same blood group varies over a wide range. Indeed, the difference, in this respect, among individuals of the same group seems to be enormous, when contrasted with the usually narrow physiological range of variation of most of the other normal components of the blood. (Thus, the isoagglutinins resemble the immune bodies in the plasma.) The fact, then, that a small part of the population normally possesses high titers of isoagglutinin has led, in the past, to the search for high-titered sera to be used as blood grouping materials. However, the difficulty of securing these sera in sufficient amounts has led to further developments in the production of isoagglutinins. Among the later procedures, are included the preparation of immune hemagglutinins by the injection of animals, and the preparation of immune isoagglutinins by the injection of human subjects. Each of these methods offers special advantages in some respects, and, particularly, to some laboratories. However, by the fractionation of normal human plasma obtained from the blood of random donors, large quantities of the anti-A and anti-B reagents may be prepared economically, and a series of other products of clinical value is simultaneously obtained.

The isoagglutinins have, so far, not been chemically separated from one another, and the group A and B specific substances present in some plasma neutralize the antibodies. It is necessary, therefore, to segregate the plasma specimens into pools which will be processed into anti-A

preparations, on the one hand, and into anti-B reagents, on the other. Thus, all of the plasmas from group A donors are pooled and processed into anti-B material. It is recognized that most of the anti-B antibody in such a pool is contributed by a relatively small fraction of the donors, and that random variation in the incidence of high-titered donors does occur, but the size of the pools, representing hundreds of donors of group A in the commercial practice, insures against much variation among different pools. The most important requirement for these pools is that they be capable of yielding satisfactory final products following fractionation, and it is, in fact, found that the methods of chemical fractionation, shortly to be discussed, are adequate for the concentration of the anti-B antibody present in random pools from group A donors. Since these donors form about 41% of the population of the United States,[7] the amount of material obtained from a given number of blood donations is large.

Formerly, the production of anti-A reagents started in a manner analogous to that just described, *i.e.*, from group-specific plasma obtained from B donors. However, it was found that such anti-A preparations were often deficient in their reactions with A_2B cells. Selection of the specimens to be included in the pools led to a reduction of the amount of reagent that could be prepared. Since group B donors form only about 10% of the population, the selection program limited the amount, not only of anti-A reagent, but also of anti-B, that could profitably be prepared from a given number of blood donations, since the two reagents are needed in the same amount in routine blood group examinations.

The problem was solved by the use of mixed blood of groups O and B, as the source of the anti-A isoagglutinins.[8] Pooled group O plasma was found to have better activity, for cells having the A_2-agglutinogen, than has pooled B plasma. That the two are approximately equal in their activity for A_1 cells was also found, and this fact might have been derived from the data in the literature.[9] Thus, O plasma would appear to have a higher concentration of the species of anti-A antibody which reacts with all group A cells, while B plasma would seem to contain more anti-A_1 agglutinin, specific for subgroup A_1 cells.[10]

Group O donors form about 45% of the population, and, by mixing blood of this group with that obtained from B donors, who constitute about 10% of the population, the amount of anti-A material that may now be produced from a given number of blood donations exceeds the amount of anti-B. (It might prove feasible to equalize production by the processing of group O plus A material.) The procedure we recom-

mend is to mix the group O and B bloods, in the same proportion as their relative incidences in the population, and to stir the mixture in the cold for about one hour prior to centrifugation. During this period, most of the anti-B antibody, present in the O plasma, is absorbed by the B cells or is neutralized by the B substance present in B plasma. By this process, group O plasma becomes the main source of the anti-A antibodies, and the B bloods make possible the removal of most of the anti-B agglutinin. Thus, the procedure could be modified to start with O plasma and B blood or cells.

A small part of the original anti-B activity remains in the O plus B plasma, after the cells are removed by centrifugation, and this antibody must be eliminated later in the process. It is undesirable to attempt complete absorption at the plasma stage, not only because the higher temperatures or the longer times needed to accomplish this would be harmful to the other products obtained by fractionation of the plasma, but also because there is a decrease of anti-A antibody, during absorption of the anti-B. This decrease of anti-A antibody is not so rapid as to be very troublesome in practice, yet it is of some theoretical interest. The phenomenon could be explained by the presence, in O plasma, of agglutinin molecules capable of reacting with either the A or the B agglutinogen (the α- + β-agglutinin of earlier workers,[10] recently called anti-C[11]) but an explanation based on the observation of "secondary binding" of heterologous agglutinins by specifically-agglutinated erythrocytes, reported by Thomsen,[12] may be in better agreement with the kinetics of the phenomenon.

Experience has shown that it is most important that all group A and AB bloods be excluded from the pool of group O and B bloods. The presence, during the absorption process, of even a small number of cells containing the A-agglutinogen will lead to a considerable loss of anti-A antibody. In the case of the anti-B isoagglutinin, also, it is important that blood containing the B-agglutinogen be absent from the bloods used, because plasma for fractionation is commonly prepared by continuous centrifugation, a process which allows partial mixing of the bloods prior to removal of the cells, and because of the presence of B substance in the plasma of B and AB donors. Thus, if successful products are to be made, errors in grouping the bloods to be included in the pools must be eliminated. In order to accomplish this, it is desirable to use potent reagents in the blood-grouping tests performed at the processing plant, but personal errors may still be expected. Accordingly, we recommend that a double blood grouping test be performed. We have found it convenient, first to test the cells of all of the

donors (cell-grouping), and then to check the sera of the donors (serum-grouping), using cells of known groups as the reagents for the latter test. By the careful use of such a procedure, pools containing several hundred blood donations may be prepared with small risk of error.

Anti-Rh Isoagglutinins

In the case of the production of anti-Rh reagents, the problems are different in many respects. It has become increasingly evident that the need for this material could not be met, unless new methods were applied. Previously, the only anti-Rh agglutinins available were obtained either by immunization of guinea pigs or rabbits, or by a painstaking search for human serum possessing sufficient antibody to be useful. The experimental sera produced by animal injection were often weak, and generally could not be used by inexperienced workers. Human sera containing large enough amounts of the antibodies are quite rare, as may be judged from the following results.

In reviewing 6000 obstetrical cases in and about Boston, during one year, we found the average incidence of *Erythroblastosis fetalis,* an evidence of sensitization of the Rh negative mother and the development of isoagglutinins in her serum, to be about one in 200 deliveries. Thus, of this group of 6000 women, only 30 might be expected to possess the agglutinins, and of these only ten had demonstrable anti-Rh antibodies, as determined by the ordinary test tube compatibility test using saline suspensions of Rh positive cells. Of the ten, four had an agglutinin titer sufficiently high to make the serum safe for routine testing after the dilution that is incurred by neutralization of the anti-A and anti-B isoagglutinins. Of the four possible donors, one was in too poor health to be bled, and one had developed agglutinin of the anti-Rh′ variety, a kind of antibody which agglutinates the cells of only 73% of the general population, and, therefore, cannot be used for the ordinary routine determination of Rh factor. The remaining two sera were of the 85% and 87% varieties, and were, therefore, generally useful. In summary, out of 6000 obstetrical cases examined, only two donors were found whose sera could be used for routine laboratory tests. Even if a liter of serum could be obtained from each of these subjects in the course of a year, this would permit of only about 20,000 tests, a number far below the requirement, even for a single city. It was, thus, imperative to find methods for using other sources of anti-Rh agglutinins.

While only a few donors were found whose sera were of high enough titer to be used directly as anti-Rh typing materials, between five and

eight times as many subjects were found whose sera contained the agglutinins in low titer. We therefore decided to collect such material and, by applying fractionation methods, to attempt the production of suitable reagents.

In order to supply a program of the desired scale, it was necessary to organize a means of collecting the anti-Rh-containing sera, obtained chiefly from those who, either by transfusion or, more commonly, by pregnancy, had been sensitized and had developed the antibodies Thus, a large number of medical centers throughout the country were visited, and arrangements were made for specimens to be sent to a central laboratory in Boston, where the presence, or absence, of anti-Rh agglutinins is determined. Larger donations of blood are then obtained from those individuals who have useful amounts of anti-Rh antibody, and, thus, substantial amounts of serum are now being collected.

Early in this study, it was found that, of the specimens submitted from men who had shown hemolytic transfusion reactions, presumably due to the Rh factor, and from women who had delivered infants with definite *Erythroblastosis fetalis*, fewer than 50% showed anti-Rh agglutinins by the ordinary saline cell suspension test tube technique. More than half of the specimens either showed no agglutinins at all, or the titers were so low as to be inconsistent with the severity of the symptoms observed. By a slide test method using whole blood, described below, many samples of serum, which would previously have been discarded as useless, are now identified as containing useful agglutinins.

Before the serum obtained by the collection program can be conveniently used for fractionation, it is necessary to prepare pools of the material. Unfortunately, it was found that, in certain cases, the process of mixing the sera resulted in a loss of potency. Thus, it was noted that addition of a 73% (or anti-Rh′) serum to an 85% or 87% (anti-Rh_0 or anti-Rh_0') serum resulted in a mixture which showed the reactions of the anti-Rh′ agglutinin: *i.e.*, it would react with cells of only 73% of the general population, and Rh_2 cells would not be agglutinated. As shown by the work of Race and of Wiener,[13, 14, 15] the phenomenon was caused by the presence, in the 73% serum (and in many other anti-Rh sera), of substances called "blocking antibodies" or "inhibitors" which interfere with the demonstration of Rh positive cells in a saline suspension. By segregating the sera collected into different pools, each containing anti-Rh agglutinins of the same specificity, a first step was made toward overcoming the difficulty.

B. FRACTIONATION PROCESS

Plasma contains a wide variety of protein constituents, and these can be separated into fractions by processes that depend upon the solubilities of the respective components, under specified conditions of pH, ionic strength, ethanol concentration, and temperature. The pro-

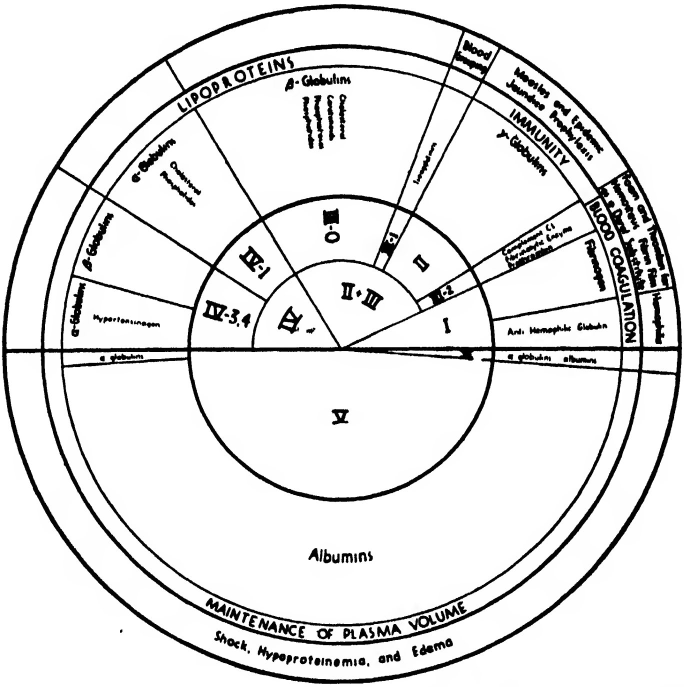


FIGURE 1. Plasma proteins, their natural functions, clinical uses, and separation into fractions. (Revised from an earlier diagram in Science **101**: 54. 1945.)

cedures used for the separation of the protein fractions, and the characterization of the fractions so obtained, are described in a series of papers.[16, 17] The approximate amounts of the various fractions obtained from normal plasma, and their principal components and uses, are illustrated in the circular graph (FIGURE 1).

In addition to the solubility properties inherent in each of the components from which the isoagglutinins are to be separated, there are several factors which govern the choice of the fractionation method:

1. The method should be compatible with the maximum use of each of the other protein constituents of the plasma.
2. The method should provide a concentration of isoagglutinins in the final solution at least eight-fold that of plasma, and this should be attained at protein concentrations in the final solution of less than 10%, since more concentrated globulin solutions possess viscosities that are undesirably high and that increase rapidly as drying occurs on the slide.
3. The method should provide reagents which are stable for considerable lengths of time. For this reason, dried products are advantageous. In order that the isoagglutinin-containing fraction may be dried, it is necessary that the lipid and carotenoid content be fairly low, or it is necessary to add stabilizing substances to protect the labile components.
4. The method should provide reagents having fibrinogen and thrombin contents so low that clots do not form, when dilute whole blood is used for agglutination tests.

It was shown by Boyd* that the isoagglutinin activity of plasma was concentrated in Fraction II + III. This fraction is precipitated under conditions of pH near 7, ionic strength about 0.09, and ethanol concentration between 20 and 25%. In the presence of such ethanol concentrations, it is necessary to work at a temperature of —5° C., or lower, in order to prevent the denaturation of some proteins. The Fraction II + III obtained under these conditions is known to contain, in addition to the isoagglutinins, most of the antibodies found in plasma, the C′1 component of complement, a pro-fibrinolytic enzyme, prothrombin, a large part of the lipoproteins, nearly all of the carotenoid pigments, some remaining fibrinogen, and other components. The further separation of the isoagglutinins from these constituents depends upon a series of observations concerning the solubility properties of the components of major interest:

1. The isoagglutinins in normal plasma have a minimum solubility at about pH 6.3. At this pH, the solubility is very low in aqueous systems at low ionic strength, and, in the presence of salt, can be held at a low value by the use of even low concentrations of ethanol. Thus far, no differences have been found between the solubilities of the anti-A, anti-B, or anti-Rh agglutinins, respectively.

* Reported in a summary of the concentration of antibodies in globulin fraction derived from human plasma by J. F. Enders.[14]

2. The larger part of the γ-globulins are least soluble in the pH range around 7.2. At pH 5.2 to 5.4, they are very soluble, however, and are, perhaps, the most soluble component of the plasma, at that pH, at fairly low ionic strengths.

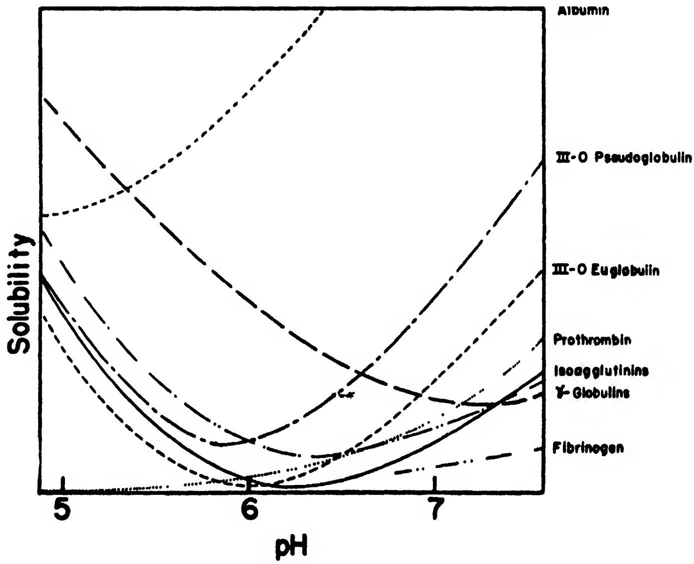


FIGURE 2. Estimated solubility behavior of some human plasma proteins

3. Prothrombin and whatever fibrinogen may not have been removed from the fraction are the least soluble components of plasma, at pH 5.2 to 5.4. Prothrombin has a very low solubility, at this pH, even in solutions of fairly high ionic strength.
4. A large part of the protein carrying lipid and carotenoid pigments is soluble at pH 7.2 to 7.6, even in the presence of fairly high concentrations of ethanol. By keeping the ionic strength low, under these conditions, most of the other components of Fraction II + III can be kept insoluble. It would appear that this lipid-containing fraction, possibly the "X-protein" of McFarlane,[19, 20] is characterized by a low density, and that this property can sometimes be used to effect the separation of the fraction by differential settling or sedimentation rates.

These observations, combined with other data, have been presented in the form of estimated solubility curves, obtained as the pH is varied

in systems of fairly low ionic strength and ethanol concentration (FIGURE 2). These curves are only intended to convey a qualitative picture of the solubility of the components of Fraction II + III. Most of the lipid present in this fraction is concentrated in the III-O euglobulin component. Two curves are presented for the behavior of the γ-globulins, one with a minimum solubility between pH 7.0 and 7.6, and another with a minimum near pH 6.3. The isoagglutinins have a solubility minimum near pH 6.3, and they are "salted-in" by small concentrations of electrolyte at this pH. These solubility relationships are the basis for most of the separations involved in the fractionation of the isoagglutinins. By the application of these principles, it has been possible to achieve satisfactory concentrations of anti-A, anti-B, and anti-Rh isoagglutinins.

C. ABSORPTION AND NEUTRALIZATION

It is essential that the products resulting from the fractionation procedure be specific for the groups of cells with which they are supposed to react. In the case of anti-B preparations, derived from group-specific plasma pools, it is only necessary to avoid contamination with material of other groups, in order to achieve specificity. Anti-A preparations, made from pools of group O plus B blood, are reabsorbed with group B cells. This is conveniently done in the course of the fractionation procedure. For example, one of the intermediate fractions may be reconstituted to neutral, isotonic solution, in a volume about one-fifth that of the original plasma, and treated with about 5 cc. of washed group B cells per liter of original plasma. The small amount of B cells, forming about 3% of the amount used for the initial absorption, has been found to be sufficient for the removal of residual anti-B antibody, when the reabsorption is allowed to proceed for about 12 hours in the cold. The product is rendered specific by this treatment and, so far as tests indicate, the reabsorption is not accompanied by loss of anti-A antibody. Thus, the reabsorption appears to be unlike the original absorption, carried out with plasma and a far higher concentration of B cells. In the case of anti-Rh preparations, the Witebsky soluble group A and B substances are added to the plasma in amounts more than sufficient to neutralize the anti-A and anti-B agglutinins present, and the products are then specific for the Rh agglutinogens.

D. STANDARDIZATION OF PRODUCTS

The large scale production of isoagglutinins permits control of the potency of the products, since the size of each preparation makes even

an elaborate testing procedure economically feasible. Uniformity from lot to lot is achieved by using large plasma pools, by blending products of different potencies, and by varying the protein concentrations in the final preparations. Standardization of the reagents is important, since they must provide clear-cut and reproducible tests in the hands of many users, and so the elimination of materials possessing low potency or showing certain other undesirable properties is necessary. In the case of the anti-A and the anti-B isoagglutinins, considerable uniformity has been achieved. Standardization of anti-Rh preparations is also important, and this presents certain special problems.

Isoagglutinin assays and chemical analyses have been our principal procedures for controlling the final products. The isoagglutinin tests made on each preparation include determinations of titer, avidity, and specificity. High avidity and a reasonably high titer are required to provide suitable reactivity, especially with weakly reactive subgroup cells, and to assure the maintenance of sufficient potency, even if some loss occurs in the dry powder over long periods of time. Complete specificity is, of course, essential. Assays for these properties have been routinely performed in our own, and in several cooperating, laboratories, by a group of investigators that includes Mr. J. W. Cameron, Dr. W. C. Boyd, Dr. L. K. Diamond, Dr. E. L. DeGowin, Capt. L. R. Newhouser, Lt. Comdr. S. T. Gibson, and Miss J. Sullivan.

Thrombin, prothrombin, and fibrinogen concentrations are determined, in order to show that the product will not cause clot formation, when it is used on the slide. When high concentrations of some of these materials make the use of a preparation objectionable, we have found that the addition of heparin to the reagent prevents clotting, even when whole blood is used on the slide. The moisture content of the dried powder is determined, and the reconstituted solution is analyzed for its protein and lipid concentrations, and for its pH and electrophoretic pattern.

Anti-A and Anti-B Isoagglutinins

To facilitate the standardization of the anti-A and anti-B reagents, two isoagglutinin preparations, one of each group, have been adopted[6] as reference standards. Each of these serves as the basis for comparison with any product of the corresponding group. Samples of the reference standards, packaged in small units as dried powders, have been made available to each of the evaluating laboratories. When unknown material is being assayed, a sample of the reference standard is dissolved and tested simultaneously, and the results of both tests are then

compared. By such a procedure, many systematic errors otherwise inherent in the assay are controlled, and, furthermore, it is possible to compare results obtained by somewhat different techniques.

The methods of isoagglutinin assay used in the various cooperating laboratories differ from one another in some details. Those adopted at the Harvard Medical School will be described briefly. By a titration method, we determine the highest dilution at which agglutination is macroscopically visible. Two-fold, serial dilutions of the antibody are prepared in test tubes, and then an equal volume of a suspension of fresh cells of the appropriate group, previously twice-washed in about fifteen volumes of saline and made up to a 2% suspension, is added to each tube. After the tubes have been well mixed by shaking, they are briefly centrifuged. The relative degrees of agglutination are then observed, as each tube is shaken gently. It is convenient to express the results in terms of the "titer index," which we define as the negative logarithm, to the base 2, of the highest dilution showing agglutination. An advantage of this method of expressing titration results is that the precision is easily stated. When the test is based on the usual method of determining the end-point, the precision is about ± 1 titer index unit. However, recently, we have adopted a mathematical device which permits maximum use of the data obtained in the titration. By this device, the titer index of an unknown preparation being compared with the reference standard is determined, not by a comparison of the respective end-points alone, but by the comparative degrees of agglutination observed in all of the tubes showing agglutination. Thus, while the technique of the titration remains unchanged, the precision of the measurement is generally improved to ± 0.5 titer index unit, or better.

Avidity tests are made on microscope slides observed with the unaided eye. A measured amount of the reference standard is mixed with the same amount of a 10% suspension of twice-washed cells. Simultaneously, on the same slide, the preparation being evaluated is treated in the same manner. As the slide is rocked, observations of three kinds are made: the time required for "first visible" agglutination, the time for an advanced stage which we have called "complete clumping," and the "size" of final clumps obtained. Each of these properties has been arbitrarily defined, in terms of the behaviors of the reference standards reacting with the cells of certain persons. Anti-A preparations are tested with A_1 cells and with cells having the A_2 agglutinogen (actually, A_2B cells are used by preference), and anti-B preparations are tested with B cells. Thus, there are, in fact, three in-

dependent sets of standard values for the speeds and sizes of agglutination.

Reproducibility of results with a given preparation, when avidity tests are performed by the same worker with cells from the same donor, is highly satisfactory. Any departure from the standard conditions, however, is liable to alter the results. The findings of different evaluators, using different cells and different techniques, have been considerably more variable, yet the results are in close enough agreement to make acceptable preparations distinguishable from unacceptable ones. It is important that the avidity of the preparations be satisfactory, since the speed at which routine blood group examinations may be performed by the slide technique depends on this property.

The products are also tested for their specificity, by a technique which is like that used for the avidity tests, except that heterologous cells are used and very long times are allowed.

High stability of the reagents in the final package is essential to their usefulness The properties of the products should remain unchanged for the longest possible period of time. Our tests have shown the anti-A and anti-B isoagglutinins to be quite stable, when their moisture content is low (under 2%), and they can be heated at 50° C. for several weeks, with only slight losses in titer and avidity. At room temperature, no appreciable change in titer or avidity has been observed. Preparations with higher moisture content are usually less stable. We have prepared special, experimental preparations of very low stability. A sample of such an anti-B preparation, kept in the dried form, but still containing much moisture, lost a considerable part of its activity in the course of a few weeks at room temperature. We found it possible to improve the activity of the degenerated product, by adding Fraction III-O, rich in lipoproteins, to the material. This experiment was undertaken because of the observation, made over a year ago, that the lipids play an important role in the phenomenon of isoagglutination by the anti-A and anti-B antibodies. It was observed that, by some lipid-extraction techniques, it was possible to decrease the isoagglutinin titer of a preparation considerably, and, furthermore, it was shown that a large part of the titer was restored, by recombining the extracted lipids with the protein. (However, the agglutination observed in the latter case was atypical, in that the clumps in the test tube broke up, or "disaggregated" on standing.) It appears that, in the case of the experimental product of low stability referred to above, the loss of activity was probably due to the instability of the lipids.

Anti-Rh Isoagglutinins

No reference standard anti-Rh preparation has yet been adopted, although it now seems desirable to establish such a standard in the future. Both "85%" and "87%" reagents are occasionally prepared, and at present the requirements for acceptance of a product must be described in absolute terms. In the test tube titration method, performed with fresh cells known to be of types Rh_0, Rh′, and Rh″, we require that the appropriate macroscopic agglutinations (*i.e.*, positive with Rh_0 alone for an 85% reagent, and with Rh_0 and Rh′ for an 87% reagent) be obtained at a 1:16 dilution, after incubation for 15 minutes at 37° C. In the avidity tests, the product is judged acceptable if it shows beginning agglutination of fresh, oxalated, whole bloods of the appropriate specifications in 30 seconds, and complete agglutination of these bloods in 120 seconds, with large final clumps. For this test, a large drop of whole blood is mixed on a slide with a smaller drop of antibody solution. The slide is rocked and observed over a source of light. The detailed technique has been described.[21] The reagents are also tested against cells of the mixed specificities more commonly encountered, and are further required to show reasonable activity against older cell samples, which ordinarily are much less sensitive than fresh ones, especially when tested on the slide. The products are also tested to prove the absence of anti-A and anti-B isoagglutinins, and any other non-specific agglutinins which might be present.

The stability of dried anti-Rh preparations is a problem for continued investigation. While some preparations of uncontrolled moisture content have shown considerable loss of activity, in the course of a few months storage, one lot of material, containing less than 1% moisture, has remained nearly unchanged for about six months. Both bovine and human serum albumin in high concentrations have recently been shown to increase the stability and potency of anti-Rh materials.[22] Albumin is particularly useful in conjunction with preparations high in anti-Rh "blocking" antibodies. Like plasma, over which it shows certain advantages, it permits the demonstration, both on the slide and in the test tube, of agglutinins which may fail to react if a saline diluent alone is used.

Anti-Rh typing reagents are now made to contain 20% of bovine serum albumin. Blood typing is performed in such a manner that the final mixtures contain between 10 and 20% of serum albumin. This use of serum albumin has usually eliminated trouble caused by anti-Rh "blocking" antibodies, and also appears to increase the stability of the

reagents. Unfortunately, the use of high concentrations of albumin increases the likelihood of non-specific reactions, and makes the neutralization of traces of anti-A and anti-B isoagglutinin more difficult, so that reagents must be carefully checked for such complications.*

SUMMARY

The principles underlying the methods for large-scale production of isoagglutinin reagents are discussed:

1. Procedures for the collection and pooling of blood donations are described. The bloods of group A donors are used for the preparation of the anti-B isoagglutinin; the bloods of group O plus B donors, for the preparation of the anti-A isoagglutinin. Blood for the production of anti-Rh reagents must, of course, be collected from donors who, as a result either of transfusion or, more commonly, of pregnancy, had developed the antibodies. By applying the fractionation procedure to plasma from random donors of the proper groups, it is possible to prepare effective anti-A and anti-B reagents, and by applying these procedures to plasma containing low titers of the anti-Rh isoagglutinin, it is possible to use the blood of between five and eight times as many donors as could be effectively used in the form of unconcentrated serum.
2. The chemical methods employed in the fractionation and concentration of the isoagglutinins make possible the use of many of the other protein components of the plasma, thus allowing a more effective utilization of the human proteins, and minimizing the cost of production of these reagents. The methods provide a fraction containing isoagglutinins concentrated at least eightfold over plasma, thus assuring the effectiveness of the blood grouping material.
3. Procedures are outlined for treating the isoagglutinin preparations, to make them specific for the groups of cells with which they are supposed to react.
4. Standardization of the reagents is important, since they must provide clear-cut and reproducible tests in the hands of many users. To assure their uniformity, certain chemical tests and immunological procedures based on the use of reference standards are routinely employed in testing the reagents.

* Note added in proof.

BIBLIOGRAPHY

1. **Cohn, E. J., J. L. Oncley, L. E. Strong, W. L. Hughes, Jr., & S. H. Armstrong, Jr.**
   1944. J. Clin. Invest. **23**: 417.
2. **Cohn, E. J.**
   1944. Am. Philosoph. Soc. **88**: 159.
3. **Cohn, E. J.**
   1945. Science **101**: 51.
4. **Pillemer, L.**
   1943. Science **97**: 75.
5. **Pillemer, L., J. L. Oncley, M. Melin, J. Elliott, & M. C. Hutchinson**
   1944. J. Clin. Invest. **23**: 550.
6. **De Gowin, E. L.**
   1944. J. Clin. Invest. **23**: 554.
7. **Wiener, A. S.**
   1943. Blood Groups and Transfusion; 304 (from data of **Snyder**). (Third ed.) Thomas. Springfield, Ill.
8. **Melin, M.**
   1945. J. Clin. Invest. **24**: 662.
9. **Bryce, L. M., & R. Jakabowicz**
   1941. Med. J. Australia **1941**(I): 290.
10. **Landsteiner, K., & D. H. Witt**
    1926. J. Immunol. **11**: 221.
11. **Wiener, A. S., & H. E. Karowe**
    1944. J. Immunol. **49**: 51.
12. **Thomsen, O.**
    1931. Z. f. Immunitätsf. **70**: 140.
13. **Race, R. R.**
    1944. Nature **153**: 771.
14. **Wiener, A. S.**
    1944. Proc. Soc. Exp. Biol. & Med. **56**: 175.
15. **Diamond, L. K., & N. M. Abelson**
    1945. J. Clin. Invest. **24**: 122.
16. **Cohn, E. J., J. A. Luetscher, Jr., J. L. Oncley, S. H. Armstrong, Jr., & B. D. Davis**
    1940. J. Am. Chem. Soc. **62**: 3396.
17. **Cohn, E. J., L. E. Strong, W. L. Hughes, Jr., J. N. Ashworth, D. J. Mulford, & H. L. Taylor**
    1946. J. Am. Chem. Soc. **68**: 459.
18. **Enders, J. F.**
    1944. J. Clin. Invest. **23**: 510.
19. **McFarlane, A. S.**
    1935. Biochem. J. **29**: 497.
20. **Pedersen, K. O.**
    1945. Ultracentrifugal Studies of Serum and Serum Fractions. Almquist & Wiksells, AB. Upsala, Sweden.
21. **Diamond, L. K., & N. M. Abelson**
    1945. J. Lab. & Clin. Med. **30**: 204.
22. **Cameron, J. W., & L. K. Diamond**
    1945. J. Clin. Invest. **24**: 793.

ISOHEMAGGLUTININ TITER AND AVIDITY*

By Louis Pillemer

Institute of Pathology, Western Reserve University, Cleveland, Ohio

INTRODUCTION

Studies by numerous authors on the concentration, characterization, and the activities of blood-grouping sera and globulins have shown that these reagents exhibit marked variations between the test-tube titer, which measures the quantity of isohemagglutinin, and the avidity* expressed as the time required for macroscopic slide agglutination. The present paper is concerned with a study of the factors responsible for these variations. Certain observations which are presented suggest that two or more serum components are involved in isohemagglutination: *i.e.*, the isohemagglutinin, *per se*, and a lipo-protein complex necessary for satisfactory avidity.

Grouping Globulins

The advent of mass blood grouping in the armed forces necessitated procurement of grouping reagents which would exhibit rapid and specific macroscopic agglutination. It became necessary, therefore, to investigate the possibilities of enhancing the avidity of grouping sera. This has been accomplished by concentrating the grouping globulins by neutral salt precipitation[1] or by methanol[2] or ethanol[3] fractionation. Table 1 summarizes the relationship between isohemagglutinin titers and the avidities of normal, pooled, group-specific sera, and of grouping globulins prepared from these sera by methanol and ethanol fractionation. The previously described methods for the determination of rate of macroscopic slide activity (avidity)[3] and of the test-tube titer by centrifugation were employed in all assays conducted in this study. Inspection of table 1 reveals: (1) that sera of equal titer may vary considerably in avidity; (2) that a serum of low titer may be more avid than a serum of relatively high titer; (3) that processing of serum into grouping globulin, by methanol or ethanol techniques, enhances the avidity as well as the test-tube titer; and (4) that, while further refinement of grouping globulin increases the titer, such proc-

* Although the term "avidity" may be vague and perhaps misleading, it is retained here, because of its general use by most investigators.

TABLE 1

THE RELATIONSHIP BETWEEN TEST TUBE TITER AND SLIDE AVIDITY IN GROUP SPECIFIC PLASMAS AND GLOBULINS

Pool or subject number	Test Tube Titer: Concentration over plasma	Test Tube Titer: Original plasma	Test Tube Titer: After concentration or immunization	Slide Avidity (Seconds): Original plasma, Initial	Slide Avidity (Seconds): Original plasma, Complete	Slide Avidity (Seconds): After concentration or immunization, Initial	Slide Avidity (Seconds): After concentration or immunization, Complete
Methanol Fractionation							
1 (Anti-A)	4	64	256	15	120	5	15
2 (Anti-A)	4	32	128	15	90	5	15
4 (Anti-A)	4	64	256	5	30	5	15
3 (Anti-B)	4	32	128	10	120	15	30
5 (Anti-B)	4	32	128	30	150	15	60
6 (Anti-B)	4	128	512	10	60	5	15
Ethanol Fractionation							
9 (Anti-A)	4	128	512	—	—	5	25
16 (Anti-A)	4	256	1024	—	—	5	15
x-2 (Anti-B)	4	128	512	15	180	5	30
88 (Anti-A)	8	64	512	15	90	5	30
90 (Anti-A)	16	128	2048	15	90	5	30
94 (Anti-A)	8	128	1024	30	90	5	15
104 (Anti-A)	16	128	2048	15	75	5	30
84 (Anti-B)	8	32	256	15	90	10	30
91 (Anti-B)	8	128	1024	30	100	5	15
89 (Anti-B)	16	64	1024	15	150	5	30

essing does not always result in an increased avidity and may, in fact, often lead to a diminished avidity.

It was noted early, by Dr. John Elliott and the author, that a fraction of plasma, which was composed of most of the serum gamma and beta globulins, prepared by a single precipitation with methanol or ethanol, was more avid than preparations which had been subjected to further purification. However, in order to prepare grouping globulins that could be dried from the frozen state without the addition of sucrose or other adjuvants, it was important to obtain fractions relatively free of lipid material. Nevertheless, as seen in TABLE 1, such additional purification made it necessary to prepare refined grouping globulin with four times the isohemagglutinin titer of the less refined fractions in order to obtain the same avidity. Even then, the size of the clumps of agglutinated red cells obtained with the more refined globulins was often much smaller and more difficult to read macro-

scopically. This suggested that, in the course of the extended purification, a factor of serum was removed which determined avidity but not titer.

Isoimmune Serums

Immunization of humans with soluble A or B substances[4] markedly increases the isohemagglutinin titers and, to a lesser extent, the avidities of their sera. In an extensive immunization program, carried out at the Army Medical School, it was noted that the avidities of the isoimmune sera vary considerably and are not closely correlated with titers. This is shown in TABLE 2 which summarizes the relationship

TABLE 2

THE RELATIONSHIP BETWEEN TEST TUBE TITER AND SLIDE AVIDITY OF SERA FROM ISOIMMUNIZED SUBJECTS*

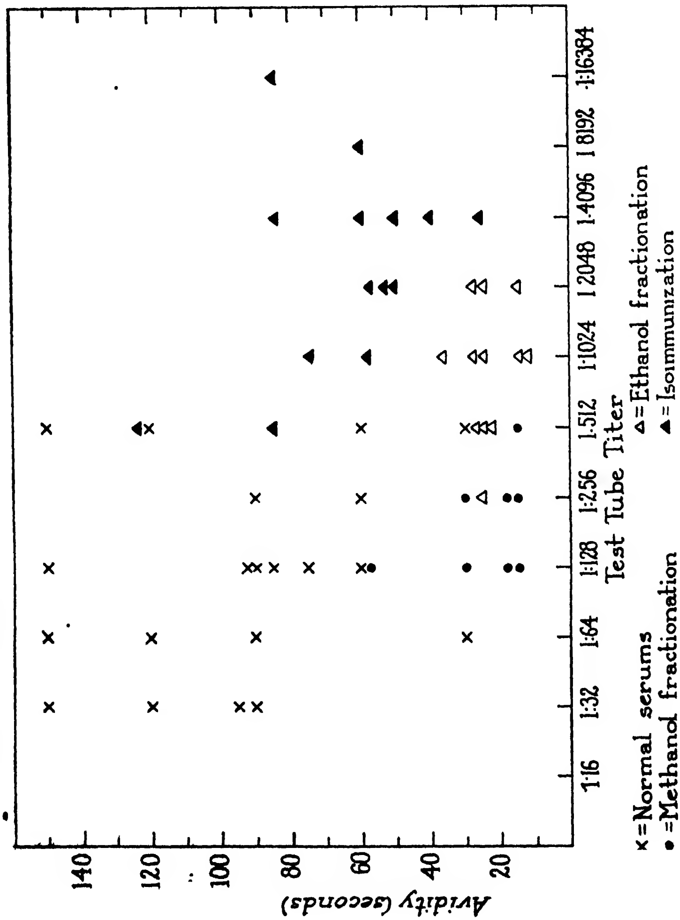
Test Tube Titer			Slide Avidity (Seconds)			
Subject number	Original plasma	After immunization	Original plasma		After immunization	
			Initial	Complete	Initial	Complete
22 (Anti-A)	1024	4096	—	—		30
12 (Anti-A)	64	2048	—	—		50
13 (Anti-A)	32	2048	—	—	7	50
16 (Anti-A)	256	4096	—	—	10	90
15 (Anti-A)	256	512	—	—	6	85
17 (Anti-A)	512	1024	—	—	10	75
14 (Anti-A)	128	4096	—	—	5	50
5 (Anti-B)	16	8192	—	—	6	60
3 (Anti-B)	512	2048	—	—	8	60
4 (Anti-B)	512	16384	—	—	15	85
7 (Anti-B)	512	4096	—	—	5	45
10 (Anti-B)	128	1024	—	—	10	60
18 (Anti-B)	1024	4096	—	—	10	60
9 (Anti-B)	64	512	—	—	15	120

* Subjects received, intravenously, 0.5 ml. soluble A and B substances (Lilly). Blood was collected between the tenth and twentieth day after immunization. The highest titers obtained are recorded in the table.

between the isohemagglutinin titers and avidities of isoimmune sera. It will be noted that artificial stimulation of the isohemagglutinin titer by injection of soluble A or B substances does not necessarily cause a proportionate increase in the serum avidity. In fact, in these experiments, it was not possible to obtain isoimmune sera of which the avidities compared favorably with those exhibited by the products of methanol or ethanol fractionation, even though the titers of the isoimmune sera were greater.

CHART I summarizes the relationship between the avidities and titers of normal sera, of grouping globulins prepared by methanol and ethanol fractionation, and of isoimmune sera. It is seen that grouping globulins, especially those prepared with methanol, show strikingly the greatest avidities, and when note is taken of their relatively low titers, as compared with isoimmune sera and with the more refined globulin preparations produced by ethanol fractionation, it again becomes evident that, during methanol fractionation, a constituent of serum not

Chart I. Relationship of the Test Tube Titer to the Avidity of Normal Serums, Isoimmune Serums and Grouping Globulins.



necessarily associated with titer was concentrated. In the course of extensive purification, this principle may be removed or destroyed, while, in isoimmunization, this material may not be present in the necessary state or in adequate amount for high avidity.

Since grouping sera produced by artificial immunization may often show a disproportionate increase of titer as compared with avidity, it is also possible that these isoagglutinins may have physicochemical characteristics different from the natural isohemagglutinins which are electrophetically beta-globulins,* and, as judged in the ultracentrifuge, protein complexes having a sedimentation constant of 20S.[5] In contrast to these natural isohemagglutinins, the induced isohemagglutinins may resemble other human antibodies by being associated with the more soluble gamma pseudoglobulins, which have sedimentation constants of about 7S. This possibility, which is under investigation, may account in part for their relatively low avidities.

Influence of Extrinsic Factors

In a series of studies on the factors responsible for the variations between test-tube titer and avidity, non-specific factors were first investigated. It was found that, at temperatures between 2° C. and 37° C., the relationship of test-tube titer to avidity was unaffected. It was further noted that hydrogen ion concentrations between pH 5.5 and 8.6 also had no influence on this relationship. Grouping sera dialyzed against 0.9% saline retained the previous titer-avidity relationship.

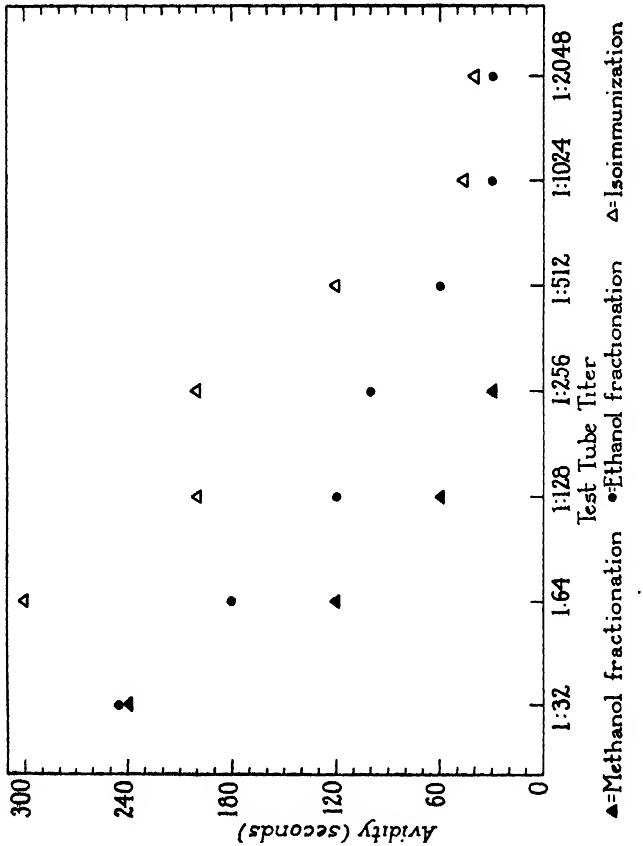
Dilution of an individual grouping reagent markedly influenced the relationship of titer to avidity. This is shown in CHART II, which graphically portrays the effect of dilution on the activities of grouping globulins prepared with methanol, with ethanol, and an isoimmune serum. This chart shows that an isohemagglutinin globulin, prepared by ethanol fractionation, and an isoimmune serum of equal titer had approximately equal avidities. A grouping globulin, prepared by methanol fractionation, although having a titer 1/8 that of the above reagents, had equal avidity. Upon dilution, the grouping reagents rapidly lost their avidities. The isoimmune sera lose this property more rapidly upon dilution than do the grouping globulins. This may be due to the concurrent concentration, during chemical fractionation, of the isoagglutinins and the other serum constituents necessary for avidity.

---

* Personal communication by Dr. J. W. Williams.

In a personal communication with Dr. Elliott, the author has learned that the addition of 2% salt solution to grouping globulins dried from the frozen state increased the avidity of the grouping reagents as compared to reconstitution of the same reagents with distilled water or 0.9% saline. We have confirmed and extended Elliott's findings.

Chart II. Influence of Dilution on the Relationship of Test Tube Titer to Avidity in Grouping Globulins and Isoimmune Serum.



The phenomenon is, perhaps, due to a disturbance of the avidity factor during drying, since we have noted that NaCl concentrations between 0.5% and 4% did not appear to influence the relationship to test-tube titer avidity of undried grouping globulins or sera.

Influence of Lipid Extraction

Since factors such as hydrogen ion concentration, temperature, or dialysis do not markedly influence the relationship of titer to avidity, the role of intrinsic factors was next investigated. The studies of Horsfall and Goodner[6] showed that extraction of anti-pneumococcal rabbit or horse serum, with organic solvents at low temperatures under very carefully controlled conditions, resulted in a loss of the agglutination power of the antiserum, with little or no loss of the protective ability. Apparently, the de-fatted anti-pneumococcal antibody combined with the pneumococcus or its specific carbohydrate without the occurrence of any visible manifestations, such as agglutination or precipitation. Horsfall and Goodner considered that the phenomenon was due either to removal of a lipid fraction which was an integral portion of the antibody molecule, or to the removal of substances present in serum which are non-specific, but are necessary for visible agglutination or precipitation.

In the present studies, both anti-A and anti-B globulins were extracted by the methods of Horsfall and Goodner, employing both Method A (alcohol and ether extraction) and Method B (alcohol, petroleum ether, and ether extraction). As observed by Horsfall and Goodner, maintenance of certain definite experimental conditions is necessary to avoid denaturation of the serum proteins and also to extract the lipids adequately.

The de-fatted grouping globulins gave clear solutions in 0.9% saline at pH 7.2. It was observed that the extracted euglobulins exhibited new physico-chemical characteristics, in that they had lost most of their euglobulin character, becoming, for the most part, soluble in water.

The effect of the extraction of grouping globulin with organic solvents on the test-tube titer and avidity is shown in TABLE 3, and it is seen

TABLE 3

EFFECT OF EXTRACTION OF GROUPING GLOBULIN WITH ORGANIC SOLVENTS

	Original Globulin		After Extraction: Method A		After Extraction: Method B		Addition of Extracted Lipids	
	Test-tube titer	Slide avidity (seconds)	Test-tube titer	Slide avidity (seconds)	Test-tube titer	Slide avidity (seconds)	Test-tube titer	Slide avidity (seconds)
104B	1024	30	256	∞	256	∞	512	150
9193A	512	30	256	∞	256	∞	256	180

that treatment by either Method A or Method B resulted in a complete loss of avidity with some loss in titer. The addition of extracted lipids, in the manner advocated by Horsfall and Goodner, resulted in partial restoration of diminished test-tube titer and in a recovery of avidity.

On the basis of these results, it is suggested that lipoidal substances or complexes may play a role in the avidity of isohemagglutinin globulins. The loss of avidity and partial loss of titer may conceivably be due to the fact that the removal of lipids sufficiently changes the physico-chemical equilibrium of the system, so as to prevent secondary antigen-antibody reactions from taking place after union has occurred.

On the other hand, the isohemagglutinin antibody or "avidity factor" may be a lipo-protein complex. The recent, important work of Pedersen[5] reveals that the isohemagglutinins, *per se*, are associated with a serum protein having a sedimentation constant of 20S, and that isohemagglutinin activity may influence the presence or absence of an X-protein complex. Lipid extraction dissociates the X-protein. The characteristics of the X-protein as described by Pedersen, and those of the avidity factor as determined by the present author, are presented in TABLE 4. It is apparent that the X-protein and the factor

TABLE 4

CERTAIN PROPERTIES OF PEDERSEN'S X-PROTEIN, THE "AVIDITY FACTOR," AND THE 20-COMPONENT OF SERUM

Property	X-protein*	Avidity factor	20-component†
1. Present in whole human serum	+	+	+
2. Varies in amount in different sera	+	+	+?
3. Dissociates when serum is diluted	+	+	−
4. Dissociates on lipid extraction	+	+	−
5. Adsorbs on specific red cells	+	+	+
6. Dissociates on storage	+	+	
7. Stabilized by sucrose	+	+	
8. Unstable on purification	+	+	−
9. Beta globulin	+	+	+?
10. Increases upon addition of gamma globulin	+	+?	−

* Pedersen describes the X-protein as being composed of albumin, globulin, and lipid, and having an electrophoretic mobility of a beta globulin and a sedimentation constant of 2.9 Svedbergs. He also states that it contains *no* isohemagglutinin activity.

† According to Pedersen, the 20-component is either a beta or alpha globulin of about 20 S. This fraction, Pedersen states, contains all of the isohemagglutinins.

in serum responsible for avidity are remarkably similar. The 20-component which, according to Pedersen, contains the isohemagglu-

tinins, is not too greatly affected by conditions which remove avidity. Pedersen's studies on the X-protein substantiate the view that isohemagglutinin avidity is associated with complexes of serum lipids and proteins and that the factors responsible for the titer and avidity are not closely related.

Influence of Formaldehyde

Eagle[7] has shown that treatment of anti-pneumococcal horse serum and diphtheria antitoxin, respectively, with appropriate amounts of formaldehyde results in a loss of agglutinating ability of anti-pneumococcal serum and of the flocculating properties of the antitoxin. However, the protective actions of the immune agents were not markedly altered. This indicates that primary union of the antigen and the treated antibody occurs, but that secondary *in vitro* manifestations are inhibited. Eagle also noted that the sera, after formaldehyde treatment, showed a greater solubility in distilled water. He attributed the loss of their agglutinating and flocculation properties to this increased solubility.

Since it was noted by the present author that certain solubility changes occur in grouping globulins extracted with organic solvents, the effect of formaldehyde on titer and avidity, as well as on the solubility of grouping globulins, was studied.

Varying amounts of formaldehyde were added to fixed amounts of grouping globulins and allowed to incubate for one hour at room temperature. The formaldehyde-treated globulins were then dialyzed against large volumes of 0.9% NaCl at 1° C. for 48 hours, the pH of each mixture was subsequently adjusted to 7.2, and their titers and avidities were tested. The results shown in TABLE 5 demonstrate that treatment of grouping globulin with an equal part of 37% formaldehyde destroyed both titer and avidity; treatment with an equal part of 9.2% formaldehyde destroyed the avidity, with a partial decrease of titer; with 2.2% formaldehyde, a small reduction in titer was observed; and treatment with an equal part of as little as 0.58% formaldehyde retarded macroscopic slide agglutination, with no reduction in titer. The deposit of formaldehyde-treated antibody on the red cell was less susceptible to aggregation, as evidenced by decreased avidity, and pressure packing by centrifugation was required to produce cohesion. Thus, combination of antigen and antibody occurred, while spontaneous aggregation was prevented. It is interesting to note that treatment with formaldehyde altered the solubility of the isohemagglutinins,

TABLE 5

INFLUENCE OF FORMALDEHYDE ON THE ISOHEMAGGLUTININ TITER AND AVIDITY OF GROUPING GLOBULIN

	Test-tube titer	Avidity (seconds)	Solubility at pH 6.35 $\gamma = 0.0150$
Original globulin (anti-B)	2048	30	Insoluble
Treatment with:			
1. Equal part of 37 per cent formaldehyde	4	∞	Mainly insoluble
2. Equal part of 9.2 per cent formaldehyde	128	∞	Soluble
3. Equal part of 2.3 per cent formaldehyde	256	180	Soluble
4. Equal part of 0.58 per cent formaldehyde	1024	75	Partially soluble
5. Equal part of 0.14 per cent formaldehyde	1024	45	Insoluble
6. Equal part of 0.9 per cent NaCl	1024	45	Insoluble

since they were no longer insoluble at the isoelectric point of the untreated protein.

CONCLUSION

The results of the above studies suggest that the differences encountered between the avidity and titer of various isohemagglutinin preparations may be due to differences in the quantity or in the physicochemical state of the various components in serum which make up the isohemagglutinin complex. Present evidence indicates that both a lipoprotein complex and a "heavy protein" are essential and integral parts of this agglutinating complex.

Caution should be observed in the concentration of grouping globulins to avoid conditions which may alter or remove this lipoprotein complex. It cannot be too strongly emphasized that avidity and stability are necessary properties of a grouping globulin. Attention should be directed to improving these properties of grouping reagents. Apparently, adequate titers are easily obtainable.

REFERENCES

1. Thalhimer, W., & S. A. Myron
1942. J. A. M. A. 118: 370.
2. Pillemer, L.
1943. Science 97: 75.

3. **Pillemer, L., J. L. Oncley, M. Melin, J. Elliott, & M. C. Hutchinson**
1944. J. Clin. Invest. **23**: 550.

4. **Witebsky, E., N. C. Klendshaj, & C. McNeil**
1944. Proc. Soc. Exp. Biol. & Med. **55**: 167.

5. **Pedersen, K. O.**
1945. Ultracentrifuge Studies on Serum and Serum Fractions. Almquist & Wiksells. Upsala, Sweden.

6. **Horsfall, F. L., Jr., & K. Goodner**
1935. J. Exp. Med. **62**: 485.

7. **Eagle, H.**
1938. J. Exp. Med. **67**: 495.

DISCUSSION OF THE PAPER

Dr. Elvin A. Kabat (*Columbia University, New York, N. Y.*):

I should like to raise the question of whether the term, "avidity," should be used, in blood group studies, to describe the time for agglutination to occur, using the slide technic. The term, with reference to antibodies, is used by immunologists and immunochemists to denote the existence of differences in antibodies, with respect to their capacity to combine with antigen. For example, two types of diphtheria antitoxins have been reported, with different "avidities," *i.e.*, combining capacities for toxin per milligram of antitoxin nitrogen.

There is no evidence to indicate that the time for agglutination to occur on a slide in any way measures avidity in this sense. The term, "slide agglutination time," might well be substituted for the term, "avidity." This would eliminate a source of potential confusion to immunologists and immunochemists.

Dr. William C. Boyd (*Boston University, School of Medicine, Boston, Mass.*):

In reply to Dr. Kabat, I do not think that the term "avidity" has ever acquired a meaning in immunology so precise that we can not continue using it for "speed of slide agglutination" in blood grouping, for a while longer. In particular, I am not sure that the prevailing meaning in immunology has ever been "the amount of antigen which a unit of antibody will combine with." The connotation has been rather that of firmness or speed of union, in my opinion.

THE ASSAY OF BLOOD GROUPING SERA: VARIATION IN REACTIVITY OF CELLS OF DIFFERENT INDIVIDUALS BELONGING TO GROUPS A AND AB

By William C. Boyd

Boston University, School of Medicine, Boston, Massachusetts

One of the consequences of the war has been a greatly increased demand for blood grouping materials, especially for the armed forces of the United States It has fallen to my lot to serve, along with various other workers, somewhat informally, as one of the persons in whose laboratory certain blood grouping preparations, prepared by commercial firms under contract with the United States Government, are tested and evaluated

Before the war, the demand for blood grouping materials was met, for the most part, by local preparation of the materials by the individuals or institutions who desired them. Other materials put out by various commercial houses were, however, gradually becoming available Some of these had been tested by myself and by various other workers, but our tests had mostly followed no set pattern, and any opinion which was rendered was, generally, based largely on the personal impression which each worker got from using the proposed new preparations to group a few known and unknown bloods.

After the armed forces began to ask for blood grouping materials in considerable quantity, various preparations were sent to my laboratory for assay and evaluation. The first preparations submitted were good, but later ones were not entirely satisfactory, and some of them were thought to be substandard by myself and by other workers. Comparison of the results of tests in various laboratories, compiled by DeGowin and others, revealed that the reports of various investigators were, in many cases, widely discrepant. This did not surprise me, as I regarded blood grouping as more of an art than a science, but it caused rather acute unhappiness among some of the individuals who had been in charge of procedures for the preparation of the materials, and among some of those who had prepared them. It became clear that it would be highly desirable, if it were possible, to develop a technique of testing which would yield more consistent results.

One improvement, suggested, I believe, by the Department of Phy-

sical Chemistry at Harvard Medical School, was the preparation of a "reference standard," one for anti-A, and one for anti-B agglutinins, with which each new preparation was thenceforth compared. As was expected, this reduced by a good deal the inconsistencies between different laboratories, since each worker tended to use the same technique of testing on the unknown and the standard. Thus, his results, although, perhaps, widely different in their numerical values from those of another worker, were more likely to show the same relative potency of the unknown compared with the standard. Application of statistical methods[2] also resulted in improvement, for it allowed an estimate of the error of the comparison to be made.

An additional fact, already known to those who had worked with blood groups for any considerable period, also gradually emerged into better perspective. This was, that the "titer" of sera or of other blood grouping preparations, although we have reason to suppose it enables an estimate to be made (with an error of perhaps 50 per cent) of the agglutinin content of the preparation, does not by any means tell the whole story. Another factor, somewhat vaguely termed "avidity," is also important, for of two sera of apparently equal antibody content, as judged by titration, one may be much more powerful and rapidly acting than the other, especially if the tests are carried out on open microscope slides, as was contemplated for work in the army.

A completely satisfactory method of estimating "avidity" has, perhaps, still to be worked out. In the case of blood grouping materials, at least, it has, on the whole, proved fairly satisfactory to determine, by tests on microscope slides and observing the reactions with the naked eye, the rate of agglutination. It has proved desirable to estimate two times: first, the interval which elapses after mixing, before agglutination visible to the naked eye appears; and, second, the interval which elapses before all of the cells seem to be involved in the agglutination. This latter state has, rather unsuitably, been called the "final stage" of agglutination, although the clumps usually continue to increase in size after this, and the degree of agglutination becomes greater, ending, in the most potent preparations, with all, or nearly all, the cells agglutinated into one single, large, macroscopically visible clump.

Avidity results from different laboratories still showed some discrepancies, however, and it was Dr. L. K. Diamond who discovered that these were due to the use of blood suspensions of different strength for the test. Work which Dr. Diamond and I carried out together showed that, when the strength of the suspension varied from 1 to 50

per cent, there was an optimum strength for avidity tests, when the suspensions were somewhat more than 10 per cent. (See TABLE 1, which, however, shows much less variation than would be found with most A_2 or A_2B cells.) For convenience, a cell suspension of 10 per

TABLE 1

EFFECT OF STRENGTH OF CELL SUSPENSION ON RATE OF AGGLUTINATION ON OPEN SLIDES

Density of cell suspension (per cent)	Time required for agglutination (seconds)	
	Initial	Big
50	4	20
40	5	20
25	4	16
10	5	15
5	3	15
2	5	20
1	6	30

cent was selected as standard, and the avidity results of different laboratories began to be more consistent.

One further source of variability still remains, however, and it is the primary purpose of this paper to discuss it. This is the intrinsic variation in agglutinability in the cells of different individuals of the same group.

The earliest workers[1, 3] observed that individuals of the same group varied in the ease with which their erythrocytes were agglutinated, and this was confirmed by later investigators.[2, 5, 6] Most of the difficulty with blood grouping globulins prepared by the Harvard method was with the anti-A agglutinin preparations. This was largely explained by the existence (already known) of at least two varieties of the A antigen, A_1 and A_2, the latter giving the weaker reactions with the usual anti-A sera.[4] Most of the preparations submitted were adequate to detect the A_1 antigen in unknown bloods, but some preparations were quite weak when tested against A_2 or A_2B cells. The latter provided one of the most critical tests of the material.

The earlier studies on variations in the sensitivity of the cells of individuals of the same blood group, especially of sub-groups A_2 and A_2B, had mostly been done by titration procedures, and not always with bloods which had been kept the same length of time after being taken, so that little was known of the degree of this variation among freshly-drawn bloods. The discovery, in Boston, of an A_2B individual whose

cells gave much weaker reactions than those of any other A_2B individual known to us (although not as weak as the reactions obtained with the only A_3B individual who seems to be known in the country), served to emphasize the importance of this variation. It gradually became apparent that variations in evaluation of different blood grouping preparations by different laboratories were, at least partly, due to the use of cells of different sensitivity.

At the suggestion of Dr. J. L. Oncley, I undertook a study of the normal range of variation in the sensitivity of normal bloods of equal age (one day old, once-washed cells), belonging to the various subgroups A_1, A_2, A_1B, and A_2B. From previous work, we had already reached the decision that the exact time of agglutination became less and less important, as the time required became longer. Accordingly, it was decided to use a mechanical timer, which marked off successively longer and longer intervals. Such a timer was very kindly constructed by Dr. Oncley, and, with its aid, over 100 bloods containing the A antigen were examined by me. This would be equivalent to testing more than 200 bloods taken at random. The intervals marked off by the machine are shown in TABLE 2.

TABLE 2

TIME INTERVALS MARKED BY MECHANICAL TIMER USED IN AGGLUTINATION EXPERIMENTS

Interval	Length (seconds)	Beginning time (seconds)
1	3.98	0
2	1.64	3.98
3	2.32	5.62
4	3.28	7.94
5	4.63	11.22
6	6.54	15.85
7	9.23	22.39
8	13.05	31.62
9	18.43	44.67
10	26.03	63.10
11	36.77	89.13
12	51.9	125.9
13	73.4	177.8
14	103.6	251.2
15	∞	354.8

It will be noted that, after the first interval, which is arbitrary, the times increase logarithmically, that is, each subsequent interval is greater than the preceding one by being multiplied by $\sqrt{2}$ This means that every other interval is double the next but one. Agglutina-

tion was carried out on open slides, and the intervals, during which the two stages of agglutination, "beginning" and "final," were observed, were recorded for each blood.

Each blood was tested with four different blood grouping preparations: (a) an absorbed anti-A serum which was fairly specific for A_1; (b) the "reference standard" prepared at Harvard; (c) a Harvard preparation made by absorption, according to Mr. Melin's procedure; and (d) an isoimmune serum prepared by injecting A substance into human volunteers of group B, according to Witebsky's technique.[7] Some of these, especially the last, were more potent than others, and suitable dilutions were used, so that the times of agglutination would be comparable. In an effort to keep conditions as comparable as possible, the same operator always worked with the same agglutinin preparation. The reproducibility of the determinations was checked by the introduction, unknown to us, of other samples of bloods already included in the day's batch, resulting in duplicate or triplicate determinations (TABLE 3). In plotting the final results, only averages were used in cases where one individual's cells were tested more than once.

TABLE 3

RESULTS OF TIMING AGGLUTINATION BY VARIOUS REAGENTS OF DUPLICATE OR TRIPLICATE SAMPLES (NOT IDENTIFIED UNTIL LATER) OF THE SAME INDIVIDUAL'S CELLS

Cells	Sera				
No.	Iso-im. 1:14	AI-41 1:10	104-B 1:3	Anti-A_1 straight	Sub-group of cell
1104	4, 9	5, 9	6, 8	6, 13	
1117	3, 9	5, 9	6, 8	7, 15	A_2
1124	4, 9	5, 8	6, 8	6, 12	
Average	4, 9	5, 9	6, 8	6, 13	
1106	3, 9	3, 6	4, 5	3, 7	
1114	4, 10	3, 6	4, 6	3, 6	A_1
1128	4, 9	4, 7	3, 5	3, 7	
Average	4, 9	3, 6	4, 5	3, 7	
1102	4, 9	3, 7	4, 6	2, 6	
1127	4, 8	3, 6	3, 6	3, 6	A_1
Average	4, 8	3, 6	3, 6	3, 6	
1107	5, 11	6, 15	6, 8	8, 14	
1122	5, 10	4, 9	6, 9	7, 15	A_2
Average	5, 10	5, 12	6, 9	7, 15	

The numbers separated by commas represent the time intervals (see TABLE 2) during which "initial" and "final" agglutination was observed.

The results of the tests are shown by FIGURES 1, 2, 3, and 4. The reactions of the very weak A_2B we had discovered are identified, in each graph, by the designation, "W. Be."

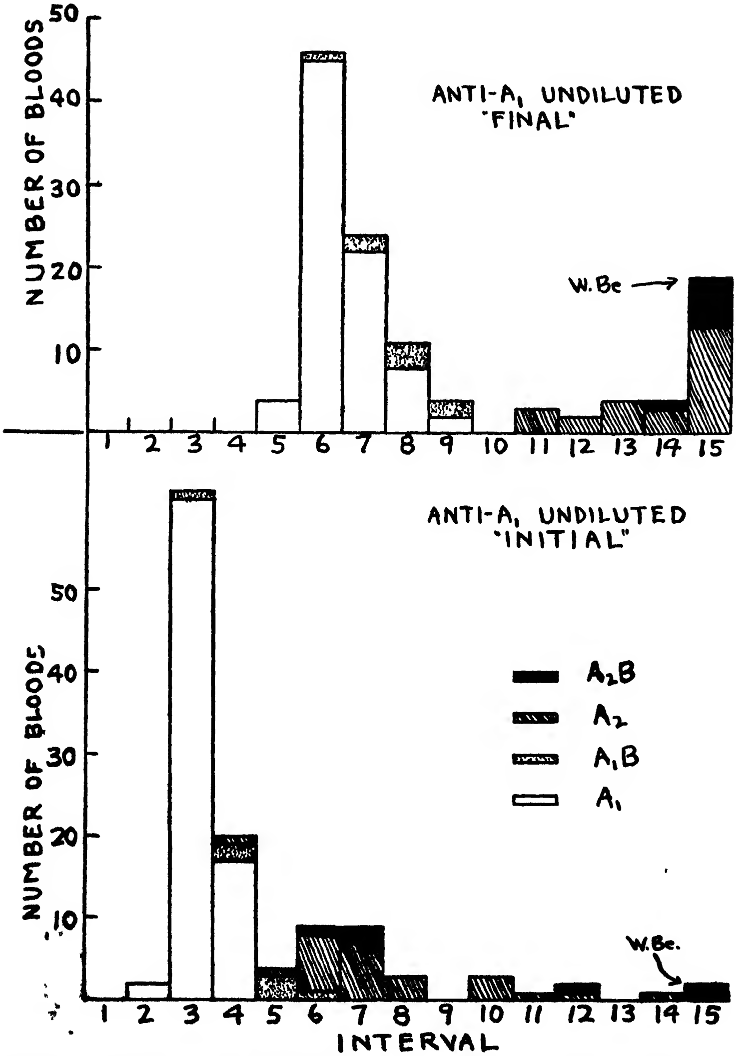


FIGURE 1. Intervals during which "initial" (below) and "final" stages of agglutination were observed with erythrocytes of various individuals, using an absorbed anti-A_1 serum.

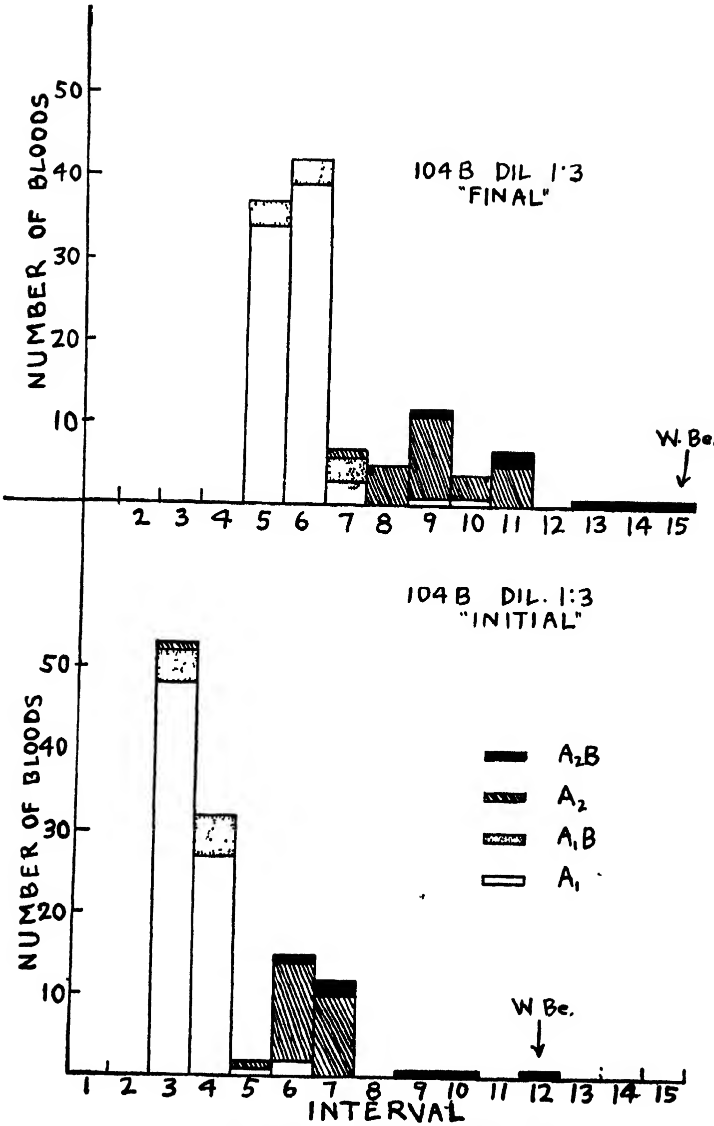


FIGURE 2. As in FIGURE 1, using the "Harvard reference standard"

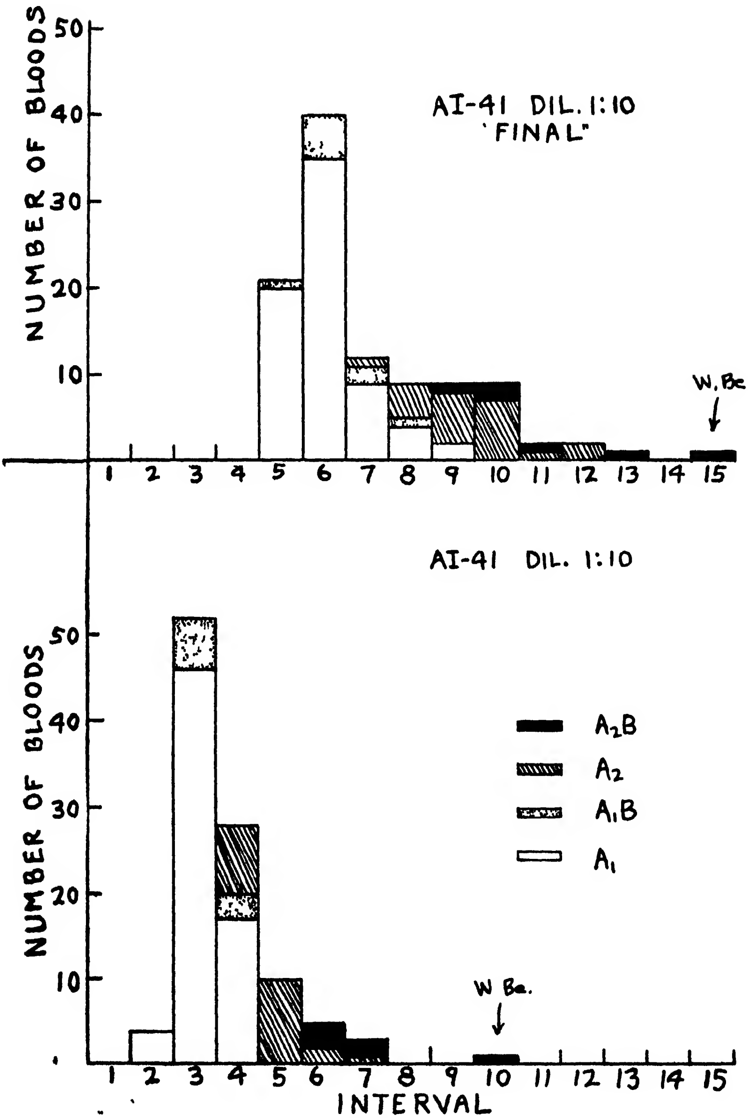


FIGURE 3. As in FIGURE 1, using a Harvard preparation made by a different method.

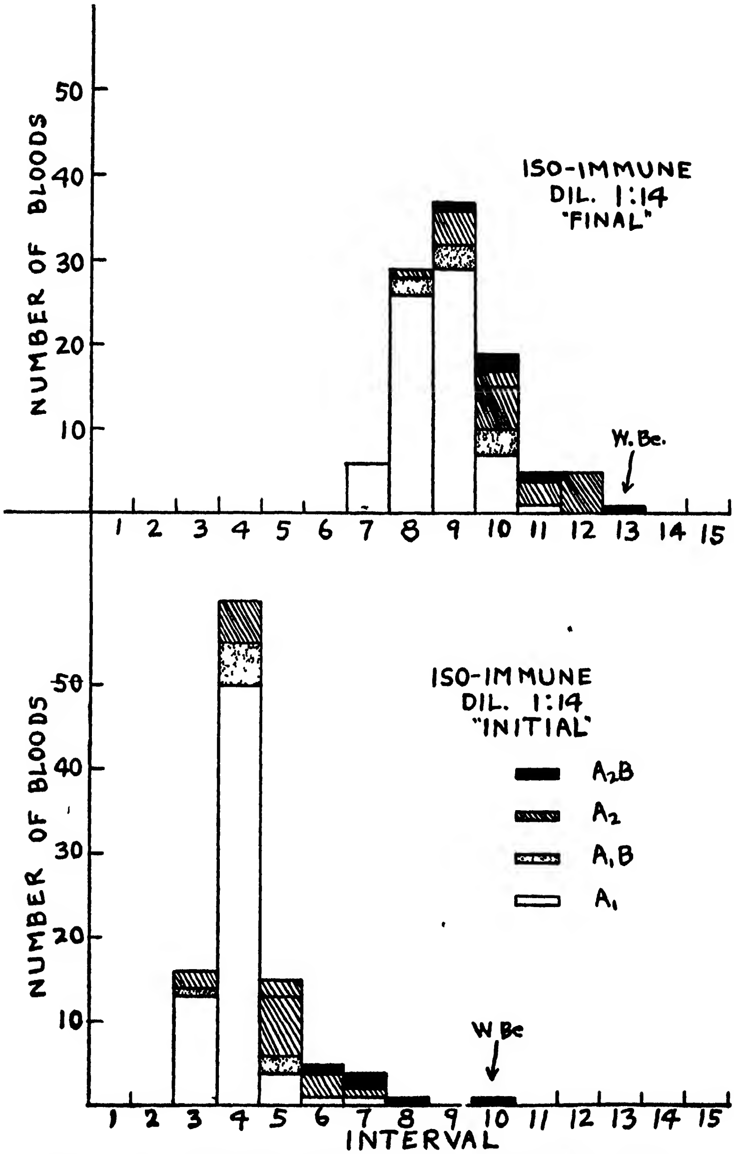


FIGURE 4. As in FIGURE 1, using an isoimmune anti-A made by Witebsky's method.

It will be seen that the difference between bloods containing A_1 and A_2 is most clearly shown by the absorbed anti-A_1 serum, as would be expected. In reading the graphs, it should be borne in mind that interval 15 is equivalent to infinity, as the reactions were not timed after this interval (354 seconds after the mixing of the serum and cells) began. The majority of the A_2B's did not show "final" agglutination before this interval began.

The Harvard "reference standard" did not differentiate so clearly between the two types of A antigen, but the graphs, nevertheless, show a bimodal type of distribution, illustrating the fact that it did not detect the A_2 antigen as well as it did the A_1. The absorbed preparation, made by Mr. Melin, did not show quite as much difference between the sub-groups, and the isoimmune sera prepared by Witebsky's method gave what was essentially a unimodal distribution, quite in line with our observations that such sera agglutinate A_2 and A_2B bloods nearly as rapidly and powerfully as they do A_1 and A_1B. For routine grouping, where one simply wants to know if the A antigen (whether strong or weak) is present, the Witebsky isoimmune sera are clearly the best.

The most novel fact which emerges from an examination of these results is the fact that the variability in A_2 bloods is nearly, if not quite, as great as the variation in A_2B bloods. Only our weakest A_2B (W. Be.) is not matched in weakness of reactivity by some A_2 blood. Also notable is the fact that bloods containing A_2 antigen vary over such a wide range in sensitivity. The moral, of course, is to use blood from individuals of *known* sensitivity in the assay of new blood grouping preparations. The most extreme A_2B bloods are probably too weak, and the strongest A_2's are probably too strong. It would, clearly, be necessary for each laboratory to find bloods of the proper degree of sensitivity, to use in testing unknown preparations. These bloods can, probably, belong either to sub-group A_2 or A_2B.

Most of the anti-B preparations which have been submitted have proved satisfactory. This is not surprising, in view of the fact that the anti-B agglutinin is generally stronger than the anti-A, and that groups B and AB, although probably divisible into sub-groups on the basis of differences in the strength of their B reactivity, never show differences in sensitivity nearly as marked as those found between A_1 and A_2, and A_1B and A_2B.

The present observations serve to emphasize the practical importance of a knowledge of the behavior of each individual serum, espe-

cially in regard to sub-groups, and the rather large (and apparently constant) individual variations in the agglutinability of erythrocytes of different individuals.

SUMMARY

In an attempt to obtain results on assay of blood grouping preparations which would give more consistent evaluation from different laboratories, it was found that one factor, among others, which caused variability in the times required for agglutination, was the intrinsic difference in sensitivity of A and AB cells from different individuals. A quantitative study of this difference (in fresh cells) has been made, and the results are presented. It was found that cells belonging to the subgroup A_2 could exhibit practically all degrees of reactivity, except that they never were as sensitive as the stronger A_1 cells, nor quite as weak as the weakest A_2B cells. In our work, the latter were available from one certain individual, whose cells seemed to give much weaker A reactions than any other A_2B's we have ever tested, but not as weak, of course, as the reaction of the cells of an A_3B individual.

REFERENCES

1. **Biffi**
   1903. Sulle emo-agglutinine del saugne umano. Ann. d'Igiene sperim. **13**: 232.
2. **Boyd, W. C.**
   1943. Fundamentals of Immunology. Interscience Publishers. New York.
3. **Landsteiner, K.**
   1901. Ueber Agglutinationserscheinungen normalen menschlichen Blutes. Wien. Klin. Woch. **14**: 1132.
   1902. Wien. Klin. Rundschau **40.**
4. **Landsteiner, K., & P. Levine**
   1929. On isoagglutinin reactions of human blood other than those defining blood·groups. J. Immunol. **17**: 1.
5. **Schiff, F., & Hübener**
   1925. Quantitative Untersuchungen uber die Empfindlichkeit menschlicher Erythrocyten für Iso-agglutinine. Z. Immunitätsf. **45**: 207.
6. **Thomsen, O., V. Friedenreich, & E. Worsaae**
   1930. Über die Möglichkeit der Existenz zweier neuer Blutgrüppen; auch ein Beitrag zur Beleuchtung sogenannter Untergruppen. Acta. path. et microbiol. Scandinav. **7**: 157.
7. **Witebsky, E., N. C. Klendshoj, & C. McNeil**
   1944. Potent typing sera produced by treatment of donors with isolated blood group specific substances. Proc. Soc. Exp. Biol. & Med. **55**: 167.

GENETIC AND CONSTITUTIONAL CAUSES OF FETAL AND NEONATAL MORBIDITY

By Philip Levine

The Ortho Research Foundation, Linden, New Jersey

Erythroblastosis fetalis, a hemolytic disease of the newborn, serves as the first example in any species of fetal and neonatal morbidity attributable to genetic and constitutional causes.[1] The fundamental mechanism is isoimmunization of the mother by hereditary serological properties in fetal blood. The blood factor most frequently involved, Rh, becomes important, theoretically and practically, by virtue of its property of antigenicity within the same species, i.e., its capacity to induce isoimmunization. Briefly stated, Rh— mothers may be immunized by their Rh+ offspring, and the hemolytic disease, resulting from the action of maternal anti-Rh agglutinins, specifically affects Rh+ , but not Rh—, offspring. Accordingly, the condition is selective, and resembles, to a striking degree, the findings in ferrets and pigs studied by Corner[2] and Robinson.[3] These workers investigated the cause of selective fetal death in a disease-free uterus, containing both normal and dead macerated fetuses, lying side by side. To make the analogy closer, one can point to twins, one of whom, Rh+, has *Erythroblastosis fetalis,* while the other, Rh—, is normal. After excluding other causes of fetal death, Corner believed that the selective fetal death may be induced by genetic and constitutional causes, and, more specifically, he cited the possibility of lethal factors. Without going into details, lethal factors described by geneticists are effective in individuals who are homozygous for a particular gene, while the erythroblastotic infant must be heterozygous. Other differentiating features will be mentioned below.

The year 1900 is significant to workers in this field, because of three important discoveries: *i.e.*, the blood groups, isoimmunization, and mendelian heredity. Although Mendel's papers first appeared in 1865, their significance was not appreciated until 35 years later, when Correns, Tchermak, and DeVries independently rediscovered mendelian laws of heredity.

In a footnote to one of his papers,[4] Landsteiner described, in 1900, the basis of his discovery of individual blood differences, *i.e.*, the action

of normal human serum on human blood cells. One year later, a fuller description of the phenomenon and its role in successful transfusions was published.[5]

Isoimmunization was discovered, in 1900, by Ehrlich and Morgenroth,[6] in experiments carried out by injection of goats with hemolyzed blood of other goats. The sera of the treated animals, initially compatible, now revealed the presence of immune antibodies (hemolysins), which differentiated numerous individuals of the same species. Isoimmunization has, since, been induced in many animal species, by the simple procedure of cross-transfusions, and the end result is the remarkably high degree of individual specificity of red blood cells. Presumably, the red blood cells of animals contain a sufficiently large number of serological properties, so that, by their permutations and combinations, each member of a species is characterized by a specific individuality of his red blood cells.

In contrast to the striking individuality in animal blood, were the findings in man, which were limited to the four blood groups and the subdivision of group A, first noted by von Dungern and Hirszfeld.[7] Fundamentally, the difficulty lay in applying successfully the phenomenon of isoimmunization in man by repeated transfusions, even at later periods, when transfusions were carried out in great numbers. Although suspected on the clinical grounds that patients receiving repeated transfusions show a higher incidence of severe reactions, the actual proof of isoimmunization—the demonstration of atypical antibodies—could not be supplied until very recently. However, another immunological procedure, heteroimmunization, yielded successful results. In 1927–1928, Landsteiner and Levine[8] discovered three new agglutinable factors, M, N, and P, which they demonstrated with the aid of rabbit sera prepared by injection of selected bloods. These discoveries served to bridge the gap between the findings in animal and human blood, and emphasized the analogy of the individuality of blood and that of skin and other tissue cells, as seen in the results of transplantation.[9] A striking illustration of these findings is seen in an experiment of Landsteiner and Levine,[10] in which the blood of each of 9 random individual workers in their laboratory could be differentiated from each other. (See TABLE 1.)

The hereditary nature of the factors A, B, M, and N, has been thoroughly established. The other properties of human blood, the factor P and the subgroups of A, are also inherited, but the differentiation of these properties is less distinct than that of A, B, M, or N.[11]

TABLE 1†

	v.d. Sch.	Ldst.	Henry	Mock	Ph. L	Hdy.	Bl.	Kr.	A.W.
Group	O	O	A	O	A	A	A	O	O
Reaction for M	+±	0	++±	++±	0	++±	++±	0	+±
Reaction for N	+	++	0	+	++±	+	0	++±	+±
Serum 1219 AA[2] (agglutinin 1[1])	0	0	+	0	0	+	0	0	0
Serum A, no. 740	+	tr	0	±	+	+	tr	++±	0
Serum B, no. 2038	±	0	*	0	*	*	*	+	0
Serum B, Mens (Ottenberg and Johnson)	0	+	*	0	*	*	*	0	±

* Isoagglutination.
† After Landsteiner & Levine.[10]

In 1940, Landsteiner and Wiener,[12] in pursuing studies on the relationship of a factor in rhesus blood related to the human property M, discovered still another which they designated as Rh. The specific anti-Rh agglutinin was found in certain anti-rhesus blood immune sera, produced in rabbits[13] or guinea pigs.[14] These workers were not aware that this factor, present in 85% of white individuals, was important clinically, and that it was identical with that described, in 1939, by Levine and Stetson.[15] Later in 1940, Wiener and Peters[16] showed that the Rh factor is antigenic in Rh negative individuals repeatedly transfused with Rh+ blood.

In 1939, Levine and Stetson described a new blood factor independent of A, B, M, N, or P, with the aid of an immune isoagglutinin, produced by a woman just delivered of a macerated fetus. This case, studied in 1937, because of a severe intra-group transfusion reaction at the first transfusion, assumes unusual importance, historically. It is the first instance in which mention is made of isoimmunization of the mother by a new blood property in fetal blood transmitted as a dominant from the father. "Presumably, the immunizing property in the blood and/or tissues of the fetus must have been inherited from the father. Since this dominant property was not present in the mother, specific immunization conceivably could occur." This blood property was present in about 80% of white individuals tested. As indicated above, Landsteiner and Wiener did not suspect that their Rh factor was identical with the blood factor of Levine and Stetson, but this was later mentioned by Wiener and Peters.* In the latter part of 1940, Levine and Katzin established that the patient, studied in 1937 with Stetson, was Rh—, and her husband Rh+, and all her compatible donors of 1937 were Rh—.†

The relationship of isoimmunization of the Rh— mother by Rh+ fetal blood to fetal and neonatal morbidity was revealed, in a series of intra-group transfusion accidents, in pregnant women not previously transfused.[17] Shortly thereafter, it became obvious that the cause of the fetal and neonatal morbidity was the condition known as *Erythroblastosis fetalis,* a hemolytic disease of the newborn. Although *Erythroblastosis fetalis* was thoroughly studied, clinically and pathologically, there was no suitable explanation for the origin of the intra-uterine blood destruction. The characteristics of the disease—early

* "In the case reported by Levine and Stetson, the incidence of bloods agglutinable by the patient's serum (based on 104 tests) was about 80%, which is not significantly different from the frequency of Rh plus blood (about 85%) . . ."

† The compatible donors were selected from the lists of the Blood Transfusion Association. All their donors were tested in April, May, and June, 1941, with the three varieties of anti-Rh agglutinins referred to in TABLE 6.

onset of jaundice, progressive anemia, enlargement of the liver and spleen, and the extra-medullary hemopoiesis seen in post-mortem examination—are the end results of intensive blood destruction. The material studied by Levine, Katzin, and Burnham[18] suggested that the maternal anti-Rh agglutinins were produced by the Rh— mother, in response to the antigenic stimulus of the foreign fetal Rh+ blood (TABLE 2).

TABLE 2

OUTCOME OF 37 PREGNANCIES IN 7 PATIENTS

Mothers	*Number*
Anti-Rh agglutinins	6
Transfused	5
Transfusion shock	5
Death after transfusion (anuria)	3
Births	*Number*
Normal babies	10
Erythroblastosis, Neonatal death, Stillbirth	15
Abortion or miscarriage	10
No data	2

It is not generally appreciated that all those findings to be reported below were made exclusively with human anti-Rh agglutinins.* Because the experimental serum was not—and still is not—satisfactory as a diagnostic reagent, Levine and co-workers[19] selected from mothers of erythroblastotic infants a large supply of several varieties of potent anti-Rh sera. Before undertaking a statistical study of mothers of erythroblastotic infants, it was necessary to standardize these anti-Rh agglutinins. At least three different anti-Rh sera were observed which gave 87%, 85%, and 73% reactions on random white individuals. The latter type of serum was independently observed by Wiener,[20] in a patient immunized by repeated transfusions.

By agreement with Landsteiner and Wiener, no attempt was made to name the agglutinins and the agglutinable properties comprising the mosaic known as Rh factor. When Levine[21] found that one of the sera contained two agglutinins, a tentative terminology was suggested for the antibodies only. Subsequently, the terminology of Wiener and Landsteiner[22] and the more inclusive terminology of Wiener[23] were applied.

* In one case, a patient of Drs Burnham, Levine, and Wiener showed that her atypical agglutinin was more or less identical, in specificity, with that of the experimental serum.

TABLE 3

AGGLUTINATION REACTIONS OF THREE ANTI-RH SERA WITH 334 RANDOM BLOODS OF ALL GROUPS*

Terminology of Wiener & Landsteiner	Incidence of type (%)	Mrs. M. F. anti-Rh_0	Mrs. M. S. anti-Rh′	Mrs. E B. anti-Rh_0'
Rh_1	71	+	+	+
Rh_2	14	+	−	+
Rh′	2	−	+	+
Rh negative	13	−	−	−
Incidence of + reactions (%)		85	73	87
Incidence of − reactions (%)		15	27	13

* The sera were obtained from mothers of erythroblastotic infants. These tests were carried out in April, May, and June, 1941.

As indicated in TABLE 3, the anti-Rh_0 serum and anti-Rh′ serum contain each one single agglutinin and the anti-Rh_0' serum contains both these antibodies. The arrangement of the reactions is such that it resembles, to a striking degree, the scheme of the four blood groups, except that the incidence of the types was distinctly different. Thus, Rh_1, which reacts with all three sera, has an incidence of 71%, as compared to 4% for group AB. It is obvious that the Rh reactions are determined by entirely different rules of heredity.

The varying importance of these sera, for the diagnosis of *Erythroblastosis fetalis* and the prevention of intra-group transfusion reactions, is indicated in TABLE 4.

TABLE 4

PERCENTAGE OF NEGATIVE REACTIONS

	Sera		
	Anti-Rh_0	Anti-Rh′	Anti-Rh_0'
Random	15	27	13
Mothers of erythroblastotic infants	92	90	90

Accordingly, it is clear that either anti-Rh_0 or anti-Rh_0' is the serum of choice, as a diagnostic reagent.[24] Historically, it is of interest that the pathogenesis of *Erythroblastosis fetalis* could have been evident, even if the anti-Rh′ serum were the only one available. The anti-Rh′ serum is not satisfactory for matings in which the father is Rh_2, and for other matings in which the mother is Rh′. Of the two sera, anti-Rh_0 and anti-Rh_0', the former is preferable, because it con-

tains a single agglutinin, and the only instances of isoimmunization which will not be detected are those in which the father is of the rare type Rh′.

A small number of Rh′ mothers could not be detected, if only anti-Rh_0′ serum were available. Actually, this type of person is immunized with greater difficulty than the Rh— individual. An analysis of the author's series of cases reveals that Rh′ mothers are four times less susceptible to immunization than are Rh— mothers.

TABLE 5

704 MOTHERS NEGATIVE WITH ANTI-RH_0

	Rh negative	Rh′
Random white*	86 7%	13 2%
Calculated Number	610	94
Observed Number	682	22
Observed %	96 9	3 1

* The values given in the first line are based on an incidence of 13% Rh negatives and 2% Rh′ In other words the total value of negative reactions with anti-Rh_0 serum is regarded as 100%

The four types of Rh, described by Levine, are now further subdivided into eight subtypes by the reactions of an additional, atypical agglutinin (anti-Rh″), which reacts on 30% of all white individuals. This agglutinin, described independently by Race[25] and Wiener,[26] reacts on two-thirds of all white individuals of type Rh_2, as defined in TABLE 3. The finer subdivision and the terminology, suggested by Wiener,[27] are given in TABLE 6, along with his values for the white population.

TABLE 6

	Reactions with			Incidence in white population
	Anti-Rh_0	Anti-Rh′	Anti-Rh″	
Rh_1Rh_2	+	+	+	16 4
Rh_2	+	0	+	12 8
Rh_1	+	+	0	54.1
Rh_0	+	0	0	2 6
Rh′Rh″	0	+	+	0 0
Rh″	0	0	+	0 3
Rh′	0	+	0	0 9
Rh negative	0	0	0	11 4

Additional terminologies, suggested by the British workers, are based on the use of letters or numbers assigned to the several genes in-

volved.[28] Consequently, the matter of terminology should be considered as tentative, and subject to later revision by an international committee of geneticists and serologists.[29]

An anti-serum produced by an Rh_1 mother should contain only one agglutinin, *i.e.*, anti-Rh″. An Rh negative individual may produce either anti-Rh_0, anti-Rh′, or anti-Rh″, or, any combination of the three agglutinins, in varying concentration, may be present in the same serum.

In general, anti-Rh″ sera are observed but rarely, and sera containing potent agglutinins of this variety alone are remarkably rare. Fortunately, the most important diagnostic serum, anti-Rh_0, is also among the most frequent. From a diagnostic and clinical viewpoint, the finer differentiation brought about by reactions of anti-Rh′ and anti-Rh″ is not very important. By and large, the clinician need not concern himself as to which of the three varieties of Rh− (Rh′, Rh′Rh″, or Rh negative) his patient belongs.[30] Consequently, he need not make any effort, at present, to commit to memory several complex terminologies, particularly since the only sera available for distribution are anti-Rh_0 and anti-Rh_0'. In this paper, the terms Rh+ and Rh− represent reactions with the standard anti-Rh_0 serum.

RACIAL INCIDENCE OF *Erythroblastosis fetalis*

Further support for the selection of anti-Rh_0 serum as the diagnostic reagent of choice was derived from an unexpected source. Potter[31] observed that *Erythroblastosis fetalis* is three times more frequent in white than in colored people, and Levine[21] showed that this was to be expected, because of a correspondingly greater incidence of Rh− individuals in the white population. This view was later extended to include Chinese[32] and Japanese.[33] These findings (TABLE 7) indicate

TABLE 7

TESTS WITH ANTI-RH_0 SERUM

Race	Number tested	+ (%)	− (%)	Incidence of *Erythroblastosis fetalis*
White[19]	334	85.0	15.0	2.1
Negro[21]	264	95.5	4.5	0.7
American Indian[34]	120	99.2	0 8	?
Chinese[32]	150	99.3	0.7	very rare
Japanese[33]	150	98 0	2.0	very rare

that the incidence of *Erythroblastosis fetalis,* in any race, is directly proportional to the incidence of Rh— individuals in any given population.

No such correlation could be demonstrated with the use of the anti-Rh′ serum (TABLE 8).

TABLE 8

TESTS WITH ANTI-RH′ SERUM

Race	Number tested	+ (%)	− (%)
White[19]	334	73	27
Negro[21]	118	46	54
American Indian[34]	69	58	42
Chinese[32]	150	93	7
Japanese[33]	150	85 4	14 6

INCOMPLETE BLOCKING OR INHIBITING ANTIBODIES

Sufficient statistical data could readily be obtained, because all women who delivered erythroblastotic infants, at any time in the past, could be included in these studies. The implication is that, at the time of the delivery, the Rh— mother must have produced anti-Rh antibodies. However, the actual demonstration of anti-Rh agglutinins was successful in only about 50% of the cases studied, within two months after the delivery of an affected infant (TABLE 9).

TABLE 9

INCIDENCE OF ANTI-RH AGGLUTININS IN 141 RH— MOTHERS*

Interval after last delivery of an affected infant	Agglutinins present	Agglutinins not found
2 months *post partum*	33	37
2 months to 1 year *post partum*	5	15
1 year or longer *post partum*	2	39
During next pregnancy	2	5
No data	0	3
Total	42	99

* Levine and associates.[19]

However, it was quite clear that even those Rh— mothers without demonstrable agglutinins were immunized, and are, therefore, subject to severe or even fatal transfusion reactions, unless Rh— blood is used.

It is now evident from the work of Race[35] and Wiener[36] that mothers

of affected infants, with no anti-Rh agglutinins, have antibodies which unite with, and specifically coat, the surface of the Rh+ red cells. However, the second stage of the reaction, the visible effect of agglutination, does not occur, under the usual conditions of testing. The phenomenon can be demonstrated by the failure of the specifically coated red cells to react with anti-Rh agglutinins. These antibodies have been termed 'incomplete,'[35] 'blocking,'[36] or 'inhibiting'[37] antibodies. At any rate, the end result *in vivo* is destruction of Rh— blood.

More recently, Diamond[38] has shown that blocking antibodies will agglutinate heavy suspensions of Rh— blood. Apparently, the essential element is the presence of normal serum, instead of normal saline as the medium for suspending the red cells (Wiener[39]).

According to Levine and Waller,[40] sera containing both blocking and agglutinating antibodies can be treated so that the former is removed and the agglutinins remain. Presumably, the blocking antibody is more readily absorbed, because non-agglutinated cells present a much larger surface area than agglutinated cells.

RH+ MOTHERS OF ERYTHROBLASTOTIC INFANTS

It is significant that a greater number of doubtful cases of *Erythroblastosis fetalis* occur in a small group of Rh+ mothers. An explanation for isoimmunization, in the smaller group of cases, the corrected value of which is about 8% instead of 10%, lies in the antigenicity of other factors, such as A and B, Hr, and finer differences of the Rh complex and, perhaps, other properties, such as that of Levine and Polayes.[41]

Perhaps the most surprising outcome of these studies is the predominant role of isoimmunization played by a particular component of only one factor. In view of the striking individuality of human and animal blood, one should have expected to find numerous blood factors, with a capacity to immunize across the placenta.

Historically, the first instance of isoimmunization by a blood factor, other than Rh, was described by Levine and Javert.[42] In this case, an Rh+ mother of an erythroblastotic infant had an atypical agglutinin which agglutinated, distinctly, the blood of her Rh— husband and her Rh+ infant. Specificity tests revealed that the agglutinin reacted strongly with all bloods failing to react with anti-Rh′ sera (Rh negative and Rh_2), and gave weak or no reactions with the remaining bloods, Rh_1 and Rh′. Parallel tests of this serum and anti-Rh′ sera revealed three types of reactions, one type reacting with both sera,

two being cross-specific types. It was significant that in no case was a blood found which failed to react with both sera. Because of the analogy to the M and N blood factors, Levine suggested that the new blood factor was allelomorphic with the factor defined by the anti-Rh′ serum. Accordingly, the letters in Rh were reversed and terms Hr and anti-Hr were selected to define the new blood factor and its corresponding antibody.[43]

Several years later, Race and Taylor[44] observed a more potent agglutinin of the same specificity. Because their serum gave more potent reactions, these workers were in a position to state that anti-Hr sera may differentiate individuals, homozygous and heterozygous for the Rh factor. As will be shown below, this view is correct, only with several limitations (Levine[24]).

It is probable that the Hr factor can explain isoimmunization in only about one-third or one-fourth of the 8% of the Rh+ group. A similar number of cases can be attributed to isoimmunization by the dominant property A and B of fetal blood, provided that the fetus belongs to the non-secretor type. Levine and Katzin,[45] and others,[46] have demonstrated that the Rh factor is not present in saliva (water-soluble form), but is present in red blood cells, presumably in an alcohol-soluble form. The antigenicity of the A and B factors in man had already been demonstrated by specific increase of isoagglutinin anti-A or anti-B, in two groups of cases: (1) transfusion of group incompatible blood, and (2) after delivery of a normal infant, most frequently the first born (Jonsson[47]).

The role of the A and B factors in inducing isoimmunization and *Erythroblastosis fetalis* can be suspected, if both parents, showing no incompatibility with any of the several anti-Rh or anti-Hr sera, have an incompatibility of the blood group factors: *i.e.*, the dominant property, A or B, in the blood of the affected infant is not present in the mother's blood.* The demonstration of specific increase of anti-A or anti-B agglutinins is confirmatory evidence. Tests of the affected infant's saliva or the deceased infant's organs should show the infant to belong to the non-secretor type.

The statistical proof for the role of A and B is given in TABLE 10. The higher value for incompatible matings in the Rh+ mothers indicates the isoimmunization by the A and B factors can explain at least some of the exceptional cases.

* In the event of an affected infant of group A, Rh+ and a group O, Rh— mother, the isoimmunization may be due to the A factor, if the Rh— mother has failed to produce antibodies for Rh.

TABLE 10

ISOIMMUNIZATION BY A AND B IN RH+ MOTHERS

Matings (white)	Compatible	Incompatible
Random	65	35
215 Rh− mothers*	75	25
28 Rh+ mothers*	50	50

* Mothers of erythroblastotic infants.
For a definition of compatible and incompatible matings, see pp 957–958

Rarely, isoimmunization, in Rh+ mothers, may be induced by finer differences in the Rh factor. From a theoretical standpoint, all individuals whose blood fails to react with any isoimmune agglutinin, of whatever specificity, could be immunized by blood sensitive to this serum. *Erythroblastosis fetalis* will result if the particular factor is limited to red blood cells. While 92% of the cases can be resolved with anti-Rh_0 serum, remarkably few cases can be attributed to isoimmunization demonstrated by anti-Rh′ or anti-Rh″ sera.

Waller, Levine, and Garrow[48] described one instance of an Rh_2 mother of an erythroblastotic infant, whose serum contained, as was to be expected, an anti-Rh′ agglutinin. In addition, another agglutinin, independent of anti-Rh″, was present, which reacted on the majority of Rh−, and some Rh_2, bloods.

In 1943, Race and Taylor[25] described a case of *Erythroblastosis fetalis*, in which the father and affected infant were of the phenotype Rh_1Rh_2, and the Rh_1 mother produced an anti-Rh″ agglutinin. An antibody of identical specificity was observed, by Levine, in an Rh_1 male patient immunized by repeated transfusions.*

In the early, exceptional, Rh+ cases, it was necessary to demonstrate atypical agglutinins, such as anti-Hr, anti-Rh′, and anti-Rh″. The remaining cases in which atypical agglutinins could not be demonstrated can now be studied, for the presence of specific blocking antibodies.

To summarize, 92% of all cases of isoimmunization can be demonstrated with the aid of the anti-Rh_0 serum. In order to demonstrate isoimmunization in the remaining cases, it is necessary to use anti-Hr, anti-Rh′, and anti-Rh″, in addition to blood grouping area. Even under these conditions, it is conceivable that, in an exceptional case, none of the sera mentioned will provide proof for isoimmunization. Consequently, the possibility must be considered that the atypical hemoly-

* The anti-Rh″ agglutinin described by Wiener[38] was produced by an Rh− mother whose serum contained, also, anti-Rh_0 agglutinin.

sin of Levine and Polayes[41] or hitherto undescribed antibodies may explain isoimmunization, in rare instances. In all Rh+ mothers of affected infants, the several criteria to establish a diagnosis in the infant should be applied most rigidly.

INCIDENCE OF *Erythroblastosis fetalis*

As indicated above, the incidence of *Erythroblastosis fetalis* depends on the number of Rh— individuals in any given population. In the New York area, the incidence of *Erythroblastosis fetalis* is given as 1:438 full term deliveries.[49] This value, however, was based exclusively on clinical and pathological findings. There is reason to believe that the condition occurs far more frequently, particularly if Rh tests are carried out in all cases of fetal and neonatal morbidity. Thus, Schwartz and Levine[50] report an incidence of at least 1:200 full term, or almost full term, deliveries.

Assuming an incidence of 1:200, this value is remarkably low, if one takes into account the 13% of all matings in which the father is Rh+ and the mother is Rh— ($85\% \times 15\% = 13\%$). There are, however, a number of factors tending to reduce this incidence, such as, for example, the current tendency to small families, and the inability of many Rh— women to respond to isoimmunization.[19] Presumably, there are genetic factors determining which of the small number of Rh— women will respond readily to isoimmunization. The nature of these genetic properties is unknown, but, certainly, they have no influence on placental permeability, as was suspected by Haldane.[51]

Although *Erythroblastosis fetalis* occurs very rarely in the first born, recent findings indicate that isoimmunization, resulting from transfusions of the Rh— female population, at any time prior to pregnancies, will tend to increase its occurrence, especially in its more severe forms (TABLE 11).

TABLE 11
Erythroblastosis fetalis IN THE FIRST BORN OF RH— WOMEN

	Transfusion history	
Severity of disease	+	−
Mild	1	4
Severe	5	4
Fetal death	10	1
Number of cases	16	9

Apparently, the immunization by previous transfusion induced an immunized state of the antibody-producing cells, so that, with the very first pregnancy, many years later, there already was sufficient intrauterine blood destruction to produce fatal forms of *Erythroblastosis fetalis*. These findings support the recommendation that no transfusion be given to young women, girls, or even female infants, unless Rh tests are carried out. Those found to be Rh— must receive Rh— blood. These precautions are necessary, also, in the administration of blood intramuscularly, since very small quantities of blood may induce isoimmunization. This simple measure, by itself, should reduce the incidence of *Erythroblastosis fetalis,* especially in its more fatal forms.[52]

FAMILIAL INCIDENCE OF *Erythroblastosis fetalis*

The obstetrical histories, given by mothers of erythroblastotic infants, fall into two groups. In certain matings, only the first one or two children are normal, while all subsequent pregnancies terminate in one of the three recognized forms of *Erythroblastosis fetalis* (*icterus gravis,* anemia, or fetal hydrops). In other matings, only one of numerous pregnancies results in fetal or neonatal morbidity, due to *Erythroblastosis fetalis.* Levine[1, 19] pointed out that the determining factor in these two contrasting family histories is the genotype of the father. If he is homozygous, all offspring must be Rh+, and every pregnancy offers an opportunity for isoimmunization of Rh— mothers. In the event of a heterozygous father, only 50% of the children will be Rh+ and the remaining 50% Rh— cannot immunize the Rh— mother. Genetic proof that the father is heterozygous can be obtained, if one of the surviving children is Rh—.

TABLE 12

SIGNIFICANT MATINGS: RH+ FATHER x RH— MOTHER

	Homozygous father	Heterozygous father
Genotypes	RhRh x rhrh	Rhrh x rhrh
Genes in Gametes	Rh rh	Rh rh rh
Offspring	Rhrh 100% Rh+	Rhrh rhrh 50% Rh+, 50% Rh—

In all matings, one, two, or more pregnancies with Rh+ fetuses are required to induce a sufficient degree of isoimmunization. However, once the Rh− mother has been immunized, all her future Rh+ fetuses will be affected with progressively more intense forms of the disease.

These remarks are based on the use of the standard diagnostic (anti-Rh_0) serum, which contains but a single antibody and will detect 92% of all mothers of affected infants. Accordingly, one may assume the very simple genetic theory indicated above, which is applicable to all except the 8% of Rh+ mothers.

From a clinical standpoint, it becomes important to differentiate, serologically, the homozygous from heterozygous father. Unfortunately, this could not be done with anti-Rh sera, but, within certain limitations, this diagnosis could be made with the anti-Hr serum of Levine. A further discussion of this subject is not possible, without taking into account the complex antigenic structure of the Rh factor and the genetics involved.

When Race and Taylor[41] stated that the blood of an individual, homozygous for the Rh factor, does not react with anti-Hr serum, they were not aware of the serologic and genetic relationship of the Hr factor and the factor described by the anti-Rh′ agglutinin, because, at this point of their studies, they had not yet found an anti-Rh′ serum. From the studies of Levine *et al.*, of 1941, it was already clear that all bloods of subtype Rh_2, whether homozygous or heterozygous, react strongly with an anti-Hr serum (TABLE 13).

TABLE 13*

Terminology of **Wiener & Landsteiner**	Mrs M. F Anti-Rh_0	Mrs M S. Anti-Rh′	Mrs. K. F. Anti-Hr.	Incidence of type (%)
Rh_1	+	+	0 or ±	71
Rh_2	+	0	+	14
Rh′	0	+	0 or ±	2
Rh negative	0	0	+	13

* Based on tests with 334 random bloods carried out in April, May, and June, 1941

It was soon pointed out that individuals of genotype Rh_1Rh_2, although heterozygous and reacting with anti-Hr serum, possess two dominant genes, Rh_1 and Rh_2, each determining the presence of a dominant blood factor capable of immunizing the Rh− mother.[24] From the point of view of fetal and neonatal morbidity, the genetically heterozygous Rh_1Rh_2 may be considered as homozygous, since, in matings with Rh− women, 100% of the offspring must be Rh+.

These findings, presented in TABLE 14 are based on reactions with anti-Rh_0, anti-Rh′, and anti-Hr sera, but the gene determining the rare property, Rh′, is not included.

TABLE 14

Phenotype	Genotype	Reactions with		
		Anti-Rh_0	Anti-Rh′	Anti-Hr
Rh_1	Rh_1Rh_1	+	+	0
	Rh_1Rh_2	+	+	±
	Rh_1rh	+	+	±
Rh_2	Rh_2Rh_2	+	0	+
	Rh_2rh	+	0	+
Rh negative	rhrh	0	0	+

A theory, describing the heredity of the Rh factor and all its variants, was proposed by Wiener,[53] and Race and Taylor.[25] According to these authors, the heredity of the Rh mosaic of several antigenic components is determined by a series of multiple alleles. Curiously enough, no provision is made for the role of a gene determining the Hr factor. In general, the experimental work in this field is hampered by the dearth of potent reagents, particularly anti-Rh″ and anti-Hr, and by the presence of genes of remarkably low frequency. By and large, it is to be expected that the final genetic theory will emerge from the statistical analysis of comprehensive studies of numerous families and racial groups, carried out with all reagents of a maximum activity.

Obviously, it is highly desirable to determine, not only which of our female population is Rh—, but also the genetic constitution of their Rh+ husbands and their surviving children. Incidentally, the bloods of all stillborn and the fetuses in miscarriages should be thoroughly tested, along with the bloods of the parents.

MECHANISM OF ISOIMMUNIZATION BY RH+ FETAL BLOOD

Since the Rh factor is not present in body fluids, it is assumed that fetal blood, in one form or another, penetrates the villus in sufficient quantity to induce isoimmunization in the mother.[19] It does not seem necessary to assume the presence of gross lesions in the placenta, which would have to recur and become operative in each succeeding

pregnancy with an Rh+ fetus, but not with an Rh− fetus. It is significant that the course of the pregnancy and the delivery of these mothers is entirely normal, in the vast majority of the cases. Although there is no direct proof that the fetal red blood cells (a large, formed element) find their way into the maternal circulation, nevertheless, the statistical data on the pathogenesis of *Erythroblastosis fetalis* permit of no other conclusion.

If one assumes that minute quantities of fetal blood, either as intact red blood cells or as stroma, pass the placental barrier, this must occur in every normal pregnancy. Isoimmunization may occur only if the fetus is Rh+ and the Rh− mother is genetically capable of producing antibodies. Accordingly, it becomes superfluous to assume the existence of genes determining placental permeability to formed elements.[51]

It is well known to the immunologist that remarkably minute amounts of antigenic material (soluble proteins, suspensions of bacteria or red blood cells) suffice to induce immunization. In recent experiments in rabbits,[54] distinct increases in agglutinin titer were observed, following 14 daily injections of 2 cc. of a 1:5000 suspension of human blood, the total volume of which was 0.0056 cc. whole blood. The corresponding value for a woman weighing 120 lbs. is only 0.13 cc.

It will be recalled that, in the latter part of the pregnancy, when isoimmunization is believed to begin, the blood vessels are adjacent to the maternal sinus, and separated from it by a single layer of cells. According to Dodds,[55] the total area of fetal villi of the human term placenta exposed to maternal sinuses is 70 sq. ft., and total length of these villi, if laid end to end, would measure 11.4 miles. One-fourth or more of the fetal blood is outside the fetus and in the placenta.

In this connection, it is significant that the pathological effects of isoimmunization by the Rh factor are observed exclusively in the fully developed, or almost fully developed, fetus. More recent data do not support the view that isoimmunization by the Rh factor *per se* plays any role in early fetal death. If there is a higher than normal incidence of miscarriages in mothers of erythroblastotic infants, this may possibly result from the effects of isoimmunization from the preceding pregnancies. At any rate, this subject merits further investigation.

Erythrocytes can be observed in the yolk sac in the four weeks old fetus,[56] and agglutinable properties could be demonstrated in the blood of the fetus, between the second and third months.[57, 58] There is reason to suspect that the more fundamental property of antigenicity, and the capacity to unite with antibodies may be inherent, even in

the forerunners of the red cells. Nevertheless, isoimmunization by the Rh factor probably is not initiated until the latter half of the pregnancy, when the blood vessels in the villi gradually approach the maternal sinuses, and are in intimate contact over an ever-increasing surface area. Pregnancy offers certain conditions which are peculiarly favorable to isoimmunization, *i.e.*, slow administration of the antigen, over a long period.

The mechanism of isoimmunization suggested is compatible with the clinical observations that, once an Rh— mother delivers an erythroblastotic infant, the condition is apt to recur, in all succeeding pregnancies, with fetuses whose red cells contain the Rh factor. Apparently, the isoimmunization is renewed, even if subsequent pregnancies are spaced at long intervals. Nevertheless, long intervals between pregnancies should be recommended, since the isoimmunization may not be as intense in the next pregnancy. In these cases, no pregnancy should be started until an interval of at least one year after all antibodies from the preceding pregnancy have disappeared

Once anti-Rh antibodies have been produced, they pass readily through the placental barrier, to act on the susceptible Rh+ fetal blood. There is no difficulty in accepting this view, since the antibodies are in solution.

In this connection, it is of interest that isoimmunization by the Rh factor must occur more frequently than is indicated by the incidence of *Erythroblastosis fetalis*. The latter condition results only from the prolonged, intra-uterine destruction of fetal blood. Consequently, one may expect to find instances of mothers with anti-Rh antibodies, but whose Rh+ infants remain normal In the several cases of this sort which have been described, it can be expected that, in the following pregnancy with an Rh+ fetus, the immunization will be sufficiently intense to produce signs and symptoms of *Erythroblastosis fetalis*. With the introduction of routine Rh tests in prenatal cases, these selected mothers should be advised regarding the spacing of later pregnancies.

Assuming the mechanism of isoimmunization described to be correct, the same considerations are applicable, also, to the several factors in fetal blood held to be responsible for isoimmunization of the Rh+ mother. Included among them, are the Hr factor, the finer differences of the Rh factor, and A and B of the non-secretor type. In this group of cases, as well as for the Rh factor, the maternal antibodies must exert their specific effect exclusively on red blood cells. Since these blood properties are not present in water-soluble form, the maternal

antibodies, after their passage into the fetal circulation, cannot be specifically inhibited, so that they are free to unite with their homologous antigens in the red blood cells.

ISOIMMUNIZATION BY THE A AND B FACTORS, AND EARLY AND LATE FETAL DEATH

Granted that a particular factor in fetal blood, Rh, may immunize the mother, with resulting intra-uterine blood destruction, the question arises, why isoimmunization induced by other factors may not also result in morbid effects on the fetus and newborn, other than *Erythroblastosis fetalis.* This question is especially pertinent, in view of the high incidence of unexplained, early and late fetal deaths and because of the remarkable individuality of human blood, resulting from the permutations and combinations of a number of well-described hereditary blood factors, in addition to Rh, such as A, B, M, N, P, and still others. The antigenicity of the factors A and B in man is already established, and their presence in water-soluble form could determine fetal morbidity, other than that due to blood destruction.

It is of interest, historically, that more than 20 years ago, incompatibility of the blood of the mother and fetus was held to be responsible for *icterus neonatorum, icterus gravis,* selective fetal death, and, also, eclampsia.[59, 60] It was also claimed by Hirszfeld[59] that, in some manner, this incompatibility adversely affected the birth weight. In this connection, the terms, *'homospecific'* and *'heterospecific'* pregnancies, were employed by Hirszfeld, and the only combinations considered by him to be compatible (homospecific) were those in which the blood group of the mother and her infant were identical. Accordingly, a mother of group A or B having an infant of group 0 was regarded by Hirszfeld as 'heterospecific.' Analysis of these early papers failed to reveal any reference to isoimmunization of the mother by dominant blood factors A and B of fetal blood, or to specific increase of the normal isoagglutinins, as direct evidence of isoimmunization.

Shortly after the pathogenesis of *Erythroblastosis fetalis* was established, Levine[1, 62] studied the bloods of women having histories of abortions and stillbirths not attributable to *Erythroblastosis fetalis.* It soon became evident that, by and large, isoimmunization by the Rh factor could be excluded, in this group of cases. However, a specific difference was observed in the blood group of the father and fetus on the one hand, and that of the mother on the other, which could be

interpreted as isoimmunization across the placenta. In these cases, the dominant properties, A and B, frequently present in the blood of the father and fetus, were lacking in the mother's blood. This sort of mating was termed by Levine, '*incompatible*,' in contrast to the '*compatible*' mating, in which the blood group of the father and mother were identical or the dominant property was present in the mother's blood.

In order to apply the same or similar statistical method, which proved so successful in the Rh studies, the two contrasting matings are classified in TABLE 15, along with their incidence.

TABLE 15
BLOOD GROUP MATINGS

Compatible	Compatible	Incompatible
♂ × ♀	♂ × ♀	♂ × ♀
0 × 0	B × AB	A × 0
0 × A	AB × AB	B × 0
0 × B		A × B
A × A		B × A
B × B		AB × 0
0 × AB		AB × A
A × AB		AB × B

If we assume random matings among the white population of the United States (group 0 = 45%, group A = 41%, group B = 10%, and group AB = 4%), then 65% of all matings are compatible and 35% are incompatible. In a group of cases, selected because of unexplained early or late fetal death, the above mentioned ratio is altered, so that there is a higher incidence of incompatible matings (TABLE 16).

TABLE 16
ISOIMMUNIZATION BY FACTORS A AND B

Matings	Compatible	Incompatible
Random	65%	35%
115 with two or more miscarriages	46	54
43 with two miscarriages or stillbirths	44	56
41 with fetal death*	41 5	58 5
215 Rh− mothers†	75	25
28 Rh+ mothers†	50	50

* Parbli[43] and Tranquilli-Leali.[44]
† Mothers of erythroblastotic infants.

As a control, a similar analysis is included of the larger group

of Rh— and the much smaller group of Rh+ mothers of erythroblastotic infants (see TABLE 10). In the Rh— series, the incidence of incompatible matings is lower than the derived value of 35%, while, in the small series of Rh+ mothers, incompatible matings are twice as frequent as in the Rh— mothers. As already mentioned, this deviation indicates that isoimmunization by the A and B factors in fetal blood, presumably of the non-secretor type, may result in *Erythroblastosis fetalis* (p. 949).

Shortly after these observations were made, Levine found reference in Taussig's book[65] to the findings of Paroli and Tranquilli-Leali, which are also recorded in TABLE 16. The values given were derived by Levine from an analysis of the data of the Italian workers, in terms of the concept of isoimmunization by fetal blood.

The data indicating a higher incidence of incompatible matings associated with abortions and stillbirths are highly suggestive and probably significant. Obviously, the statistical proof could not be expected to be as convincing for the heterogenous group of abortions and stillbirths as for the clearly defined clinical entity of *Erythroblastosis fetalis*. There are, apparently, several etiologic causes held to be responsible for fetal death, in the latter group, and these studies provide evidence that one of these causes is isoimmunization by the A and B factors of fetal blood

From a small number of cases, specific increase of the normal isoagglutinin provides additional proof for the role of isoimmunization by the fetal blood factors A and B, in abortions and stillbirths

The data on a remarkable family, with incompatibility of the Rh and A factors and selective fetal death, are presented in TABLE 17.

TABLE 17

Family F.C.B.	Group	Anti-Rh
Father	A	+
Mother	O	—
1. Child M	O	+
2. Stillborn		
3. Child C	A	—
4. Stillborn		
4. Stillborn		
5. Child L	O	—
6. Child K	O	—
7. Stillborn		
8. Stillborn		

The father must be heterozygous for both Rh and A, so that the

offspring should be 50% Rh+ and 50% Rh−, and 50% group A and 50% group O. Of the four surviving children, three have the recessive properties of the mother. Apparently, three of the four offspring having the dominant properties A and/or Rh of the father result in stillbirths.

In any random heredity study, the presence of only a few families similar to that cited above will serve to confirm the old observation of Hirszfeld on selective fetal death affecting individuals inheriting the dominant property of the father. As far back as 1924, this author found a lower incidence of group A offspring in the mating father A x mother O, than in the mating father O x mother A. Although this view was later abandoned by Hirszfeld himself, Levine revived his theory and found additional supporting evidence in an analysis of seven additional heredity studies (TABLE 18).

TABLE 18*

INCIDENCE OF GROUP A OFFSPRING IN MATINGS O x A AND A x O

	Father x Mother O x A	Father x Mother A x O
Hirszfeld (15 authors)	65 1	56 0
Hirszfeld and Hirszfeld	63 1	60 5
Landsteiner and Levine	60 6	44 8
Wiener and Vaisberg	54 8	52 5
Clausen	66 3	62 0
Vuori	46 7	63 8
Landsteiner and Wiener	68 4	25 0
Levine and Landsteiner	77 1	45 2

* From Levine.[1]

The results, in six of the seven additional heredity studies, confirm the concept of selective fetal death resulting from isoimmunization by the A factor. (In the study of Vuori, the families are made of only one or two children, in contrast to those of the other workers.) These findings suggest that, in future studies on the heredity of the several properties, it is essential to take into account a full obstetrical history, and record the outcome in all pregnancies.

It is significant that both matings, A x O or O x A, satisfy Hirszfeld's definition of heterospecific pregnancy, *i.e.*, a mother O having a child A and a mother A having a child O. Hirszfeld's observations were made in the period just before Bernstein announced his theory of the heredity of the four blood groups. It will be recalled that, according to the now discarded theory of von Dungern and Hirszfeld, an O

mother could have an AB child. Actually, Hirszfeld believed that O mothers do not have AB children, because this constituted a double heterospecific combination and he hints that the combination is lethal for these offspring. When Hirszfeld finally accepted Bernstein's view, that a group O mother could not have an AB child, because of genetic reasons, his significant observations were abandoned. In any event, it is now evident that these early observations could not be interpreted properly, in the absence of the concept of transplacental isoimmunization of the mother by the dominant hereditary factor in fetal blood.

It must be emphasized that, in this group of cases, one is dealing with the larger group of secretors of the A and B substances. Accordingly, isoimmunization may result from the passage of body fluids containing A and B substance, and manifestations of the isoimmunization may, therefore, be present in the early part of the pregnancy. Furthermore, the isoagglutinins anti-A or anti-B are normally present. Nothing is as yet known of how the reaction of the specifically increased isoagglutinin and the A and B factor in body fluids and tissues exerts its damage on the fetus. In general, one may speculate that fatal effects may result, when the reaction takes place in, or upon, vital organs or tissues.

GENERAL CONSIDERATIONS

In a sense, the data on *Erythroblastosis fetalis* tend to give additional support to the findings on the relationship of isoimmunization, by the A and B factors, to a special group of early and late fetal and neonatal morbidity. In this connection, the recent statistical studies of Gardiner and Yerushalmy, on the incidence of fetal death, are significant.[66, 67] These workers write as follows: "One may speculate that among other things, the father may also play an important part in these cases of repeated loss to the family. . . . It may therefore, be indicated that the study of infant loss should embrace also factors in the father. This seems to be especially important in the cases of habitual abortions and in cases of families in which many infants have been lost through stillbirths and neonatal morbidity."

Additional evidence for the role of isoimmunization by pregnancy is derived from the studies of Corner[2] and Robinson[3] on selective fetal death in ferrets and pigs. These workers found normal fetuses alongside of dead and macerated fetuses, in uterus which was free of detectable disease. After excluding other causes of fetal death, Corner concluded that genetic and constitutional factors, probably lethals,

may be responsible for a high incidence of fetal death. Levine believes that an alternative explanation, compatible with genetic and constitutional consideration, is isoimmunization of dominant heredity properties in fetal blood. It is a well-established fact that the phenomenon of isoimmunization by repeated transfusions is more readily demonstrated in many animal species than in man. Assuming similar types of placenta, isoimmunization by pregnancy and specific loss of fetuses can be expected to occur in any animal species in which individual blood differences are demonstrable by cross-transfusions within the species.

The effects of isoimmunization by pregnancy differ distinctly from those of genetic lethal factors. In the former, all affected fetuses must be heterozygous for the particular gene, while lethal factors affect individuals who are homozygous for a certain gene. Furthermore, the results of isoimmunization are observable either in the early or more fully developed fetus, while lethals operate over a far wider range; *i.e.*, from the fertilized ovum, the fetus, or the mature individual.

Haldane[51] states that ". . . the effects of the Rh-rh gene difference certainly account for more human deaths than any other gene difference so far known, and, very possibly, far more than all other known gene differences together." These significant remarks were made prior to Levine's findings on the role of A and B factors in causing abortions and stillbirths, other than *Erythroblastosis fetalis.* Further work is required in order to evaluate the quantitative roles of isoimmunization by the factors Rh and A and B, in inducing the several varieties of fetal and neonatal morbidity.

In all cases of fetal death due to isoimmunization by the Rh factor, there is a more rapid loss of the less frequent rh gene than of the dominant Rh gene, every affected individual being heterozygous. Hogben[68] assumes the occurrence of mutations, with loss of the Rh gene at such a rate that the several populations remain in a state of equilibrium. The objection is raised by Haldane,[69] and Fischer, Race, and Taylor,[70] that this would require an unbelievably high mutation rate. The present view is that populations are not stable, and Haldane calculates that selection, at its present rate, would reduce the frequency of Rh negative individuals, in about 600 generations, from the present value to 1%.

The data presented provide proof for the role of the Rh factor in the pathogenesis of *Erythroblastosis fetalis,* and there are indications that isoimmunization by A and B may cause a certain proportion of miscarriages and stillbirths, not due to *Erythroblastosis fetalis.* Very

early in the course of studies on isoimmunization, it was emphasized that the same statistical method could be employed to study complications of pregnancy and neonatal period, in order to determine whether or not these conditions are associated with isoimmunization by Rh, or any blood factor with capacity to immunize across the placenta.[71] Recently, Yannet[72] applied these findings to a study of an undifferentiated group of feeblemindedness in infants and children. In a study of 122 mothers of this group of children, Yannet[72] and Snyder[73] found a somewhat greater than normal incidence of Rh reactions The relationship of this group of feeblemindedness to *kernicterus* is still to be determined.

ADDENDUM

In the past year, considerable progress has been made in the improvement of methods for the direct demonstration of blocking antibodies (Diamond,[74] Wiener,[75] and Coombs, Mourant, and Race[76]). The most essential factor is the use of serum or bovine albumin, instead of saline, as the suspending medium for the Rh positive blood cells. Levine and Bernstein[77] succeeded in obtaining direct reactions of blocking antibodies and Rh positive cells, suspended in suitable concentrations of acacia and polyvinyl alcohol.

Coombs, Mourant, and Race[76] showed that the blocked-out Rh positive cells can be agglutinated by precipitins for human globulin. This technique is very useful in a group of cases characterized by persistence of the blocking antibody into subsequent pregnancies. Application of this technique makes it possible to differentiate specifically coated Rh positive cells, in which case the infant will have symptoms of blood destruction, from genetically Rh negative infants, who should remain normal. Although the same result may be obtained by resuspending the washed infant's cells in serum or bovine albumin, the technique suggested by the British workers and termed "developing test" by Hill and Haberman,[78] appears to be more sensitive.

Preliminary data indicate that the severity of the condition, in the infant, could now be more closely correlated with quantitative studies of the several antibodies, in the mother's serum,[79] but further studies are indicated.[80]

With the availability of larger supplies of human anti-Rh sera and newer methods for testing, routine Rh studies should be carried out, on a state-wide basis, in all prenatal cases, as part of a public health program.

So far as the genetics of the Rh-Hr factors are concerned, the experiment shown in TABLE 1 serves as the crossroads for divergent theories of multiple alleles (Wiener[81, 82]) and closely linked genes (Fisher and Race,[83] Levine[80]). Levine pointed out that the genetic relationship of the factors described by anti-Rh′ and anti-Hr′, which give only three types of reactions, is closer than that revealed by anti-Rh_0 and anti-Rh′ agglutinins. Accordingly, there are three closely linked genes, Rh′-Hr′, Rh_0-Hr_0, and Rh″-Hr″, each pair of which form three genotypes and three phenotypes. The corresponding terms suggested by Fisher and Race are C-c, D-d, and E-e.

BIBLIOGRAPHY

1. **Levine, P.**
   1943. J. Heredity **34**: 71.
2. **Corner, G. W.**
   1923. Am. J. Anat. **31**: 523.
3. **Robinson, A.**
   1921. Edin. Med. J. **26**: 137, 209.
4. **Landsteiner, K.**
   1900. Zentralbl. f. Bakt. **27**: 357.
5. **Landsteiner, K.**
   1901. Wien. klin. Woch. **14**: 1132.
6. **Ehrlich, P., & J. Morgenroth**
   1900. Berl. klin. Woch. **37**: 453.
7. **von Dungern, E., & L. Hirszfeld**
   1911. Ztschr. f. Immunität. **8**: 526.
8. **Landsteiner, K., & P. Levine**
   1928. J. Exp. Med. **47**: 757.
9. **Landsteiner, K.**
   1930. Nobel Prize Lecture. Science **73**: 403.
10. **Landsteiner, K., & P. Levine**
    1929. J. Immunol. **17**: 1.
11. **Landsteiner, K., & P. Levine**
    1930. J. Immunol. **18**: 87.
12. **Landsteiner, K., & A. S. Wiener**
    1940. Proc. Soc. Exp. Biol. & Med. **43**: 22.
13. **Landsteiner, K., & A. S. Wiener**
    1931. J. Exp. Med. **74**: 309.
14. **Landsteiner, K., & A. S. Wiener**
    1942. Proc. Soc. Exp. Biol. & Med. **51**: 313.
15. **Levine, P., & R. Stetson**
    1939. J. A. M. A. **113**: 126.
16. **Wiener, A. S., & H. R. Peters**
    1940. Ann. Int. Med. **13**: 2306.
17. **Levine, P., & E. M. Katzin**
    1940. Proc. Soc. Exp. Biol. & Med. **45**: 343.
18. **Levine, P., E. M. Katzin, & L. Burnham**
    1941. J. A. M. A. **116**: 825.
19. **Levine, P., L. Burnham, E. M. Katzin, & P. Vogel**
    1941. Am. J. Obst. & Gyn. **42**: 925.

20. **Wiener, A. S.**
1941. Arch. Path. **32**: 227.
21. **Levine, P.**
1942. Science **96**: 452.
22. **Wiener, A. S., & K. Landsteiner**
1943. Proc. Soc. Exp. Biol. & Med. **54**: 167.
23. **Wiener, A. S.**
1944. Science **99**: 532.
24. **Levine, P.**
1943. J. Pediat. **23**: 656.
25. **Race, R. R., *et al.***
1943. Nature **152**: 563.
26. **Wiener, A. S., & E. Sonn**
1943. J. Immunol. **47**: 461.
27. **Wiener, A. S.**
1945. Am. J. Clin. Path. **15**: 106.
28. **Murray, J.**
1944. Nature **154**: 701.
29. **Levine, P.**
1945. Science **102**: 1.
30. **Levine, P.**
1945. Am. J. Obst. & Gyn. **49**: 810.
31. **Potter, E.**
1940. J. A. M. A. **115**: 996.
32. **Levine, P., & H. Wong**
1943. Am. J. Obst. & Gyn. **45**: 832.
33. **Waller, R. K., & P. Levine**
1944. Science **100**: 453.
34. **Landsteiner, K., A. S. Wiener, & G. A. Matson**
1942. J. Exp. Med. **76**: 73.
35. **Race, R. R.**
1944. Nature **153**: 771.
36. **Wiener, A. S.**
1944. Proc. Soc. Exp. Biol. & Med. **56**: 173.
37. **Diamond, L. K., & N. M. Abelson**
1945. J. Clin. Invest. **24**: 122.
38. **Diamond, L. K., & N. M. Abelson**
1945. J. Lab. & Clin. Med. **30**: 204.
39. **Wiener, A. S.**
Personal communication.
40. **Levine, P., & R. K. Waller**
1946. Science **103**: 389.
41. **Levine, P., & S. H. Polayes**
1941. Ann. Int. Med. **14**: 1903.
42. **Levine, P., & C. T. Javert**
Cited by **Levine *et al.***[19]
43. **Levine, P.**
1941. Yearbook Path. & Immun. **506**.
44. **Race, R. R., & G. L. Taylor**
1943. Nature **152**: 300.
45. **Levine, P., & E. M. Katzin**
1941. Proc. Soc. Exp. Biol. & Med. **47**: 215.
46. **Wiener, A. S., & S. Forer**
1941. Proc. Soc. Exp. Biol. & Med. **48**: 126.
47. **Jonsson, B.**
1936. Acta Path. et Microbiol. Scand. **13**: 424.

48. **Waller, R. K., P. Levine, & I. Garrow**
1944. Am. J. Clin. Path. **14**: 756.
49. **Javert, C. T.**
1942. Surg. Gyn. & Obst. **74**: 1.
50. **Schwartz, H., & P. Levine**
1943. Am. J. Obst. & Gyn. **46**: 827.
51. **Haldane, J. B. S.**
1942. Ann. Eugenics **11**: 333.
52. **Levine, P.**
1945. J. A. M. A. **128**: 946.
53. **Wiener, A. S.**
1944. Science **100**: 594.
54. **Levine, P.**
1944. Arch. Path. **37**: 83.
55. **Dodds, G. S.**
1922–1923. Anat. Rec. **24**: 287.
56. **Arey, L. B.**
1942. Developmental Anatomy. W. B. Saunders Co. Philadelphia.
57. **Kemp, T.**
1930. Acta. Path. et Microbiol. Scand. Supp. **7**: 62.
58. **Moreau, P.**
1935. Rev. Belge. Scien. Med. **7**.
59. **Hirszfeld, L.**
1928. Konstitutionsserologie und Blutgruppenforschung. J. Springer. Berlin.
60. **Ottenberg, R.**
1923. J. A. M. A. **81**: 295.
61. **Hirszfeld, L., & H. Zborowski**
1926. Klin. Woch. **17**: 741.
62. **Levine, P.**
1942. West. J. Surg. **50**: 468.
63. **Paroli, G.**
1928. Rivista Ital. di Ginecologia **7**: 388.
64. **Tranquilli-Leali, K.**
1932. Rivista Ital. di Ginecologia **14**: 492.
65. **Taussig, F. J.**
1936. Abortions. Spontaneous and Induced: 100. C. V. Mosby Co. St. Louis.
66. **Gardiner, E. M., & J. Yerushalmy**
1939. Am. J. Hyg. **30**: 11.
67. **Yerushalmy, J., E. M. Gardiner, & C. E. Palmer**
1941. Pub. Health Reports **56**: 1463.
68. **Hogben, L.**
1943. Nature **152**: 721.
1944. Nature **153**: 222.
69. **Haldane, J. B. S.**
1944. Nature **153**: 106.
70. **Fischer, R. A., R. R. Race, & G. D. Taylor**
1944. Nature **153**: 106.
71. **Levine, P.**
1942. Am. J. Clin. Path. **11**: 898.
72. **Yannet, H.**
1944. Bull. N. Y. Acad. Med. **20**: 562.
73. **Snyder, L. H., M. D. Schonfeld, & E. M. O. Herman**
1945. J. Heredity. **36**: 9.
74. **Diamond, L. K., & R. L. Denton**
1945. J. Lab. & Clin. Med. **30**: 821.

75. **Wiener, A. S.**
1945. J. Lab. & Clin. Med. **30:** 662.
76. **Coombs, R. R. A., A. E. Mourant, & R. R. Race**
1946. Brit. J. Exp. Path. **26:** 255.
1946. Lancet **1:** 264.
77. **Levine, P.**
In preparation.
78. **Hill, J., & S. Haberman**
1946. Texas State J. Med. **42:** 193.
79. **Wiener, A. S.**
1946. Am. J. Clin. Path. **16:** 319.
80. **Levine, P.**
Am. J. Clin. Path. (In press.)
81. **Wiener, A. S., E. B. Sonn, & H. R. Polivka**
1946. Proc. Soc. Exp. Biol. & Med. **61:** 382.
82. **Wiener, A. S.**
1946. Brit. Med. J. **1:** 982.
83. **Fisher, R. R., & R. R. Race**
1946. Nature **157:** 48.

DISCUSSION OF THE PAPER

Dr. S. H. Polayes (*Brooklyn, N. Y.*):

In regard to isoimmunization with the A (or B) factor, both Dr. Levine and Dr. Wiener are acquainted with a series of cases of *Erythroblastosis fetalis* which I have collected, in which the mothers of those babies are Rh positive. To date, I have a total of nine such instances, out of about 150 cases of *Erythroblastosis fetalis,* which I have personally encountered. The mothers of all of these nine cases are group O Rh positive, and the offspring, group A (some Rh positive, others Rh negative). In each instance. the Rh factor and all of the Rh variants known at the time the cases were being studied, as well as the Hr antigen, were excluded as possible immunizing antigens. Significantly enough, the anti-A agglutinin titer of the serum of the mothers of these infants was about three times as high as that found in fifty group O mothers of normal A children, and about five times as high as that of fifty group O nulliparas studied. Clinically, the children fulfilled all the criteria for the diagnosis of *Erythroblastosis fetalis* (hemolytic anemia of the new-born). One of the cases which I recently reported (Am. J. Diseases of Children **69:** 99. 1945) came to postmortem examination. It showed the classical anatomic changes of *Erythroblastosis fetalis,* including *kernicterus,* which is almost a pathognomonic finding in this disease. This case should answer Dr. Levine's question as to whether there are any cases of *Erythroblastosis fetalis* resulting from possible isoimmunization with the A and B factor, in which the anatomic findings of the disease have been demonstrated. These findings seem to support the contention that isoimmunization with the A (and B?) antigen may occur, resulting in hemolytic anemia of the newborn, by a mechanism similar to that already established for the Rh factor by the monumental work of Dr. Levine. It is quite likely, however, that the milder types of hemolytic anemia of the newborn may result from isoimmunization with the A (or B) factor, in contradistinction to the severer forms of the disease resulting from Rh isoimmunization.

THE Rh SERIES OF GENES, WITH SPECIAL REFERENCE TO NOMENCLATURE*

By Alexander S. Wiener and Eve B. Sonn

Serological Laboratory, Office of the Chief Medical Examiner, New York, N. Y.

Although only five years have elapsed since the Rh factor was first described by Landsteiner and Wiener,[1] this aspect of individual differences of human blood has developed into a complex subject, with important applications in clinical medicine, anthropology, and forensic medicine. In fact, the subject of the Rh factors appears to exceed, in complexity, all previously obtained knowledge concerning individual differences in human blood,[2] including the original four Landsteiner blood groups, the sub-groups of A and AB, and the M, N, and P agglutinogens of Landsteiner and Levine. In place of the original, single Rh factor, transmitted by a pair of allelic genes, *Rh* and *rh,* three Rh factors are known at the present time, together with a so-called Hr factor, and these, in combination, give rise to a large series of different varieties of Rh agglutinogens, determined by a corresponding set of at least ten or more allelic genes.

If any progress is to be made, or even if any work at all is to be possible, it is necessary to designate the Rh factor in some way. Therefore, as soon as sufficient evidence had been accumulated concerning the serologic and genetic behavior of the Rh factors, we devised a nomenclature, based on these firmly established serologic and genetic facts. Since the original designations have already been widely adopted and have proved satisfactory, because they are relatively simple to learn and work with, it would appear that no change should be made, at least for the time being, unless such a change offered a distinct advantage. Unfortunately, however, a number of other workers have recently proposed other methods of designation. As will be shown, the alternative designations that have been proposed have no special advantage, but can only give rise to confusion in an already complicated subject.

As long as the serologic and genetic behavior of the Rh factors remained unclear, the only sensible procedure was to use a temporary system of numbering for the various anti-Rh sera, and that was our procedure in our earliest publications. However, as soon as increased

* Aided by a grant from the United Hospital Fund of New York City.

knowledge justified the change, a more rational nomenclature was devised, which made use of the experimental findings. Murray's[3] recent suggestion to designate the various Rh antisera as anti-Rh_1, anti-Rh_2, anti-Rh_3, and anti-Rh_4, etc., is, therefore, a step backwards, because it does not make use of the established genetic and serologic facts. Moreover, there are $4 \times 3 \times 2 \times 1$ or 24 possible ways of numbering, so that the adoption of Murray's suggestion would be the first step in a confusion of nomenclature far worse than the Moss-Jansky travesty. (Incidentally, hardly anyone would be willing to give up the name, Rh negative, which is so well established in the medical literature, for Murray's name, Rh_{456}.)

Some workers seem to feel that every new finding concerning the Rh factors, such as the discovery of a new variety of anti-Rh serum, calls for a complete revision of the nomenclature. If one accepted this type of reasoning, an architect who planned to add a new floor to a building would have to throw over the entire superstructure, in order to build a new foundation. Actually, as will be shown, the present nomenclature is based on firmly established facts, and any new finding can readily be incorporated, by making suitable additions and revisions, without disturbing the foundation. This can best be demonstrated by reviewing the development of knowledge concerning the Rh blood factors, step by step, from the earliest observations, when only a single Rh factor was known.

The original, immune anti-rhesus serum of Landsteiner and Wiener,[4] as is well known, gave 85 per cent positive reactions on the bloods of white individuals. According to the original designation, bloods reacting with the anti-rhesus sera were said to be Rh positive, and bloods not agglutinated by these antisera were said to be Rh negative. Rh-positive blood contains an agglutinogen, Rh, absent from Rh-negative blood, and hereditarily transmitted by a pair of allelic genes, *Rh* and *rh*. The sera of Rh-negative individuals sensitized to the Rh factor most frequently contains an agglutinin giving 85 per cent positive reactions, in parallel with anti-rhesus serum. Under the present designations,[5,6] this variety of anti-Rh agglutinin is known as anti-Rh_0, and the corresponding factor, detected by it, as Rh_0. Using the present nomenclature, the two original Rh types would be designated as shown in TABLE 1.

The next step, in the evolution of knowledge of the Rh blood types, was the discovery (Wiener[7]), in a patient who had had an intragroup hemolytic transfusion reaction, of an agglutinin reacting with the

TABLE 1

Designations	Reactions with human anti-Rh_0 serum or immune animal anti-rhesus serum	Approximate frequencies among white individuals (per cent)	Genotypes
Rh+ (Rh_0+)	Positive	85	Rh_oRh_o Rh_orh
Rh− (Rh_0−)	Negative	15	*rhrh*

bloods of only 70 per cent of the white population, independent of the blood groups, M, N, or P agglutinogens. Parallel studies, on the animal immune anti-rhesus sera, proved that the antigen detected by the new antiserum was related to the Rh factor. To indicate this, the new serum was, at first, designated anti-Rh_1, but, more recently, the designation has been changed to anti-Rh′. The factor Rh′, detected by the antiserum now designated anti-Rh′, is transmitted as a simple Mendelian dominant, so that, if Rh′ were the only Rh factor known, the existence of a pair of allelic genes, *Rh′* and *rh′*, could be postulated (TABLE 2). From the onset, however, all tests with anti-Rh′

TABLE 2

Red cells	Reaction with anti-Rh′ serum	Approximate frequencies among white individuals (per cent)	Genotypes
Rh′+	Positive	70	*Rh′Rh′* *Rh′rh′*
Rh′−	Negative	30	*rh′rh′*

serum were made in parallel with the standard anti-Rh_0 serum. At first (Wiener[7]), it seemed that the combined reactions with the two sorts of anti-Rh sera gave rise to only three types, analogous to the subgroups of A in relation to group O, and the original designations used were based on this serologic and genetic similarity (TABLE 3).

If, instead of anti-Rh′, one substituted anti-A_1, and, in place of anti-Rh_0, one substituted anti-A, then type Rh_1 would correspond to subgroup A_1, type Rh_2 to subgroup A_2, while Rh negative blood would correspond to group O blood. It was soon found (Landsteiner and Wiener[4]), however, that there exist rare bloods (approximate frequency, 2 per cent), which react with anti-Rh′ serum, but not with

TABLE 3

Original designation of types	Alternative designation (present nomenclature)	Reactions with antisera		Approximate frequencies among white individuals (per cent)	Genotypes
		Anti-Rh_0	Anti-Rh′		
Rh_1	Rh_1 (Rh_0')	+	+	70	Rh_1Rh_1 Rh_1Rh_2 Rh_1rh
Rh_2	Rh_2 (Rh_0)	+	−	15	Rh_2Rh_2 Rh_2rh
Rh−	Rh−	−	−	15	*rhrh*

anti-Rh_0 serum, and, therefore, are designated as type Rh′. It was immediately obvious, from the distribution of the resulting four Rh types, that the reactions of anti-Rh_0 and anti-Rh′ are not independent, otherwise the frequency of the doubly negative bloods should equal (0.15) (0.30) = .045 or 4.5 per cent, instead of 15 per cent. Moreover, factor Rh_0 is not related to factor Rh′, as the agglutinogen A is related to agglutinogen B, genetically, otherwise the frequency of the doubly positive type Rh_1 (or Rh_0') could not exceed 50 per cent. The only remaining possibility is that Rh_0 and Rh′ are related to each other, like "partial antigens." That this is the correct interpretation, follows from the fact that, when a blood contains both factors Rh_0 and Rh′, these factors are usually transmitted together, so that the offspring inherits both factors, or neither, from the parent possessing them (Wiener and Landsteiner[8]). This indicates that, when a blood contains both the factors Rh_0 and Rh′, the factors are part of a complex agglutinogen Rh_1 (or Rh_0'), transmitted, as a unit by means of a corresponding gene Rh_1 (or Rh_0'). While the designation, Rh_0', is undoubtedly clearer, the designation, Rh_1, is to be preferred, because it is simpler. Of course, in occasional bloods, the factors Rh_0 and Rh′ will be related to each other, like the agglutinogens A and B in group AB blood, as follows from the fact that bloods exist which contain Rh_0 or Rh′ alone.

To explain the findings up to this point on a genetic basis, it is necessary to postulate the existence of four allelic genes, Rh_1 (or Rh_0'), Rh_2 (or Rh_0), *Rh′* and *rh*. Family data supporting this hypothesis have been obtained by Wiener and Landsteiner. The genetic and serologic facts concerning the factors Rh_0 and Rh′ can now be summarized, as shown in TABLE 4.

TABLE 4

Bloods containing Rh_0 (Rh_0+) 85 per cent					Bloods lacking Rh_0 (Rh_0-) 15 per cent				
Designation of types	Reactions with sera: Anti-Rh′	Reactions with sera: Anti-Rh_0	Frequencies (per cent)	Possible genotypes	Designation of types	Reactions with sera: Anti-Rh′	Reactions with sera: Anti-Rh_0	Frequencies (per cent)	Possible genotypes
Rh_1 (Rh_0')	+	+	70	Rh_1Rh_1 Rh_1Rh_2 Rh_1Rh' Rh_1rh	Rh′	+	−	2	$Rh'Rh'$ $Rh'rh$
Rh_2 (Rh_0)	−	+	15	Rh_2Rh_2 Rh_2rh	Rh−	−	−	13	$rhrh$

The original two allelic genes, *Rh* and *rh*, are related to the four allelic genes under discussion, as shown in FIGURE 1. It will be seen that the old names are retained, but applied to types (or genes) which

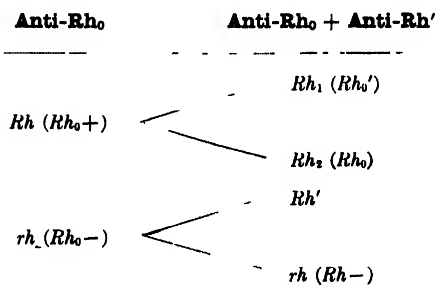


FIGURE 1. Illustrates how anti-Rh′ doubles the number of differentiable types and genes with special reference to effect on nomenclature

are only part of the original, more inclusive type, while new names are required only for blood removed from the main types. Thus, blood reacting with the standard anti-Rh_0 serum and formerly designated merely, Rh+ or Rh_0+, can now be placed in one of two types, depending on its reaction with anti-Rh′ serum. Those bloods reacting with the anti-Rh′ serum are now known as Rh_0' or, more simply, as Rh_1, while those bloods not reacting with anti-Rh′ become Rh_0 "proper" (or type Rh_2). Similarly, the original Rh negative type is subdivided into Rh negative "proper" and the rare type, Rh′.

Further complications resulted from the discovery of a third type of Rh antiserum, reacting with the bloods of only 30 per cent of white individuals.[9, 10, 11] By parallel tests with the anti-rhesus serum and using the same reasoning outlined above for anti-Rh′ serum, it was shown that the new Rh factor is related to the original Rh_0 factor, in the same way that Rh′ is related to Rh_0. Moreover, statistical and genetic studies proved that the new factor was related to Rh′, in the same way that agglutinogen B is related to agglutinogen A. Therefore, the new Rh factor is on the same plane as Rh′ and, to indicate this, it is designated as Rh″, and the corresponding agglutinin or serum as anti-Rh″.

As shown in FIGURE 2, the anti-Rh″ agglutinin again doubles the number of types of blood possible, so that, with the sera anti-Rh_0, anti-Rh′, and anti-Rh″, in combination, eight Rh types can be distinguished. Obviously, the anti-Rh″ serum doubles the number of Rh blood types possible, because each of the four types just described, Rh_1,

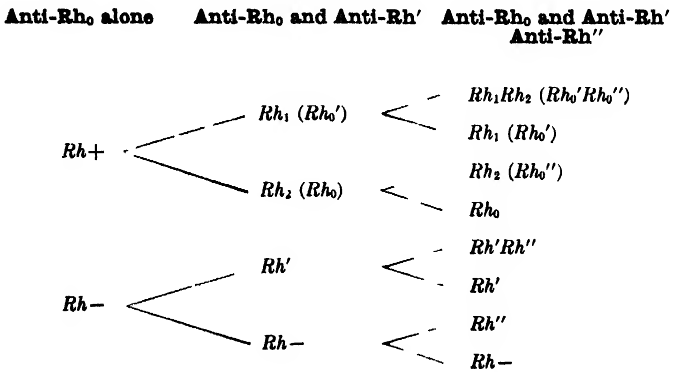


FIGURE 2. Illustrates how anti-Rh″ doubles again the number of differentiable types, and the effect on the designation of the types.

Rh_2, Rh′, and Rh negative, can be subdivided into two types, depending upon the reaction of the blood with anti-Rh″ serum. In naming these types, the principles just enunciated are merely extended, so that, again, the names of the four types detected by anti-Rh_0 and anti-Rh′ together are retained for subdivisions of these types, while new names are only required for the four additional types distinguished with the aid of the anti-Rh″ serum.

The designation, Rh_0, for the original Rh factor, was selected to

emphasize its special position in relation to Rh′ and Rh″. As has been mentioned, anti-Rh′ and anti-Rh″ sera are related to each other, like anti-A and anti-B, so that, if the reactions with anti-Rh_0 are omitted, four types can be distinguished which are serologically and genetically analogous to the four blood groups. To emphasize the special position of anti-Rh_0 serum, it is convenient to list the reactions of the eight Rh types in two columns, of four types each, as shown in TABLE 5

TABLE 5
SCHEME OF THE EIGHT RH BLOOD TYPES AND FOUR CLASSES

Classes	Bloods containing Rh_0				Bloods lacking Rh_0			
	Reactions with antisera			Types	Reactions with antisera			Types
	Anti-Rh′	Anti-Rh″	Anti-Rh_0		Anti-Rh′	Anti-Rh″	Anti-Rh_0	
W	−	−	+	Rh_0	−	−	−	Neg
U	+	−	+	Rh_1	+	−	−	Rh′
V	−	+	+	Rh_2	−	+	−	Rh″
UV	+	+	+	Rh_1Rh_2	+	+	−	Rh′Rh″

In other words, the scheme of eight Rh types resembles a double scheme of the four blood groups, the distinguishing feature between the two sets, of four types each, being the reactions with anti-Rh_0 serum. If this analogy to the four blood groups is borne in mind, it becomes a simple matter to understand and remember the scheme of the eight Rh blood types.

To account for the hereditary transmission of the eight Rh blood types, as Wiener[9, 12] first pointed out, it is necessary to postulate the existence of at least six allelic genes, instead of the four we have already described. The designations of the six genes and the reactions which they determine are given in TABLE 6. For the sake of simplicity, we shall, for the moment, disregard the rare genes, Rh_y and Rh_z (cf. TABLE 12), and the so-called intermediate genes. Before proceeding further with the nomenclature of the Rh blood types, it may be of interest to examine the available data as to the hereditary transmission of the Rh blood types, with special reference to the six-gene theory.

Under the six-gene theory, twenty-one genotypes are theoretically possible, and these correspond to the eight Rh blood types, as shown in TABLE 7. Accordingly, the following three rules of heredity should hold:[13]

TABLE 6

THE SIX STANDARD RH GENES

Genes	Reactions with Rh antisera		
	Rh′	Rh″	Rh_0
rh	−	−	
Rh_0	−	−	+
Rh′	+	−	−
Rh_1	+	−	+
Rh″	−	+	−
Rh_2	−	+	+

TABLE 7

THE EIGHT RH TYPES AND THEIR 21 GENOTYPES

Rh types	Genotypes
Neg.	*rhrh*
Rh′	*Rh′Rh′* and *Rh′rh*
Rh″	*Rh″Rh″* and *Rh″rh*
Rh′Rh″	*Rh′Rh″*
Rh_0	Rh_0Rh_0 and Rh_0rh
Rh_1	Rh_1Rh_1, Rh_1rh, Rh_1Rh', Rh_1Rh_0 and $Rh'Rh_0$
Rh_2	Rh_2Rh_2, Rh_2rh, Rh_2Rh'', Rh_2Rh_0 and $Rh''Rh_0$
Rh_1Rh_2	Rh_1Rh_2, Rh_1Rh'' and $Rh'Rh_2$

(1) The factors Rh_0, Rh′, and Rh″ are transmitted as simple Mendelian dominants and, therefore, cannot appear in the blood of a child, unless present in the blood of one or both parents. In testing this rule, it must be borne in mind that, while there are three Rh *factors*, there are five Rh *agglutinogens*, Rh_0, Rh′, Rh_1, Rh″, and Rh_2, where Rh_1 contains the factors Rh_0 and Rh′ together, and Rh_2 contains the factors Rh_0 and Rh″ together. While agglutinogens Rh_1 and Rh_2 are usually transmitted as units by the corresponding genes Rh_1 and Rh_2, in occasional individuals (genotypes Rh_0Rh' and Rh_0Rh'') the factors will exist as distinct agglutinogens that segregate genetically. Therefore, while agglutinogens Rh_1 and Rh_2 will usually behave as if they were simple Mendelian dominants, there will be rare families where the child will belong to type Rh_1 (or type Rh_2), and yet neither parent will have the agglutinogen Rh_1 (or Rh_2). For example, in the rare mating Rh_0 × Rh′, usually one-fourth of the children will give reactions corresponding to type Rh_1. In TABLE 8, are summarized the

TABLE 8

SUMMARY OF AUTHORS' FAMILY MATERIAL TO DATE

Mating	Number of families	Number of children of types							
		Rh-	Rh_1	Rh_2	Rh_1Rh_2	Rh_0	Rh′	Rh″	Totals
Neg x Neg	4	14	0	0	0	0	0	0	14
Neg x Rh_1	49	25	73	0	0	7	0	0	105
Neg x Rh_2	16	10	0	20	0	0	0	0	30
Neg x Rh_1Rh_2	15	(1)	18	14	0	0	0	0	33
Neg x Rh_0	3	3	0	0	0	3	0	0	6
Neg x Rh′	2	1	0	(1)	0	0	1	0	3
Neg x Rh″	1	1	0	0	0	0	0	5	6
Rh_1 x Rh_1	26	5	62	0	0	4	1	0	72
Rh_1 x Rh_2	21	6	15	7	18	0	2	0	48
Rh_1 x Rh_1Rh_2	27	0	46	8	25	0	0	0	79
Rh_1 x Rh_0	1	1	0	0	0	0	0	0	1
Rh_1 x Rh′	3	0	3	0	0	0	0	0	3
Rh_1 x Rh″	4	1	2	0	3	0	0	0	6
Rh_2 x Rh_2	2	1	0	4	0	0	0	0	5
Rh_2 x Rh_1Rh_2	6	0	2	8	8	0	0	0	18
Rh_2 x Rh_0	2	0	0	1	0	4	0	0	5
Rh_1Rh_2 x Rh_1Rh_2	10	0	3	1	17	0	0	0	21
Rh_1Rh_2 x Rh_0	2	0	1	1	0	0	0	0	2
Rh_1Rh_2 x Rh′	2	0	1	1	2	0	0	0	4
Rh_0 x Rh′	1	0	0	0	0	1	1	0	2
Totals	197	69	226	66	73	19	5	5	463

authors' investigations on the Rh blood types, which, to date, include 197 families with 468 children. Only a single exception to the dominance rule of heredity of the three Rh factors was encountered, namely: a family in which the mother was Rh negative, the father, type Rh′, and the child, type Rh_2. Since the father belonged to type N, and the child to type M, illegitimacy is the obvious explanation for this apparent contradiction to the genetic theory.

(2) Parents belonging to type Rh_1Rh_2 or the rare type Rh′Rh″ cannot have children belonging to type Rh_0 or Rh negative. Similarly, parents of types Rh_0 and Rh negative cannot have children of type Rh_1Rh_2 or Rh′Rh″. This law is obvious, because, under the six-gene theory, parents of types Rh_1Rh_2 and Rh′Rh″ must transmit one of the genes Rh_1, Rh_2, Rh', or Rh'' to each child, while parents of types Rh_0 and Rh negative can only transmit either an Rh_0 or rh gene to each child. As shown in TABLE 8, our investigations yielded only a single exception to the second law of heredity, and, in this case too, we obtained evidence that the child in question was illegitimate. On the basis of these results, we have been applying the Rh blood types in medico-

legal cases of disputed parentage. For reasons which will become evident, later on, we now feel that, while the first rule of heredity can be applied without reservation, a slight qualification may be necessary for exclusions of paternity based on the second rule of heredity.

(3) Further exclusions of parentage are possible, when more than one child is available for testing. For example, in the mating $Rh_1 \times$ Rh negative, if there are any Rh-negative children, then the Rh_1 parent must belong to genotype Rh_1rh, and children belonging to any type other than Rh_1 or negative cannot occur in that family. On the other hand, in families $Rh_1 \times$ Rh negative, where there are any children of types Rh_0 or Rh′, there can be no Rh-negative children. The validity of this rule cannot be checked from TABLE 8, because all families with the same kind of mating have been combined. However, if our original articles are consulted, in which each family is listed separately, it will be seen that, in our investigations, there have been no contradictions to the third rule of inheritance of the Rh blood types. On the whole, therefore, our family investigations fully support the six-gene theory of inheritance of the Rh blood types.

The other method of testing the genetic theory of the Rh blood types is to calculate the gene frequencies from the distribution of the Rh blood types in the general population and then to determine whether the sum of these calculated gene frequencies differs significantly from 100 per cent.

In order to facilitate the calculation of the gene frequencies, it should be pointed out again that, if the reactions of the anti-Rh_0 serum are omitted, the eight Rh blood types reduce to a scheme of four classes, W, U, V, and UV, analogous to the four classic blood groups, where

$$\mathrm{W} = \mathrm{Rh_0} + \mathrm{Rh\ negative} \tag{1}$$
$$\mathrm{U} = \mathrm{Rh_1} + \mathrm{Rh'} \tag{2}$$
$$\mathrm{V} = \mathrm{Rh_2} + \mathrm{Rh''} \tag{3}$$
$$\mathrm{UV} = \mathrm{Rh_1Rh_2} + \mathrm{Rh'Rh''}. \tag{4}$$

The four classes are hereditarily transmitted, like the four blood groups, by triple allelic genes, W, U, and V, where

$$W = Rh_0 + rh \tag{5}$$
$$U = Rh_1 + Rh' \tag{6}$$
$$V = Rh_2 + Rh''. \tag{7}$$

The frequencies of the six Rh genes can now be calculated with the following formulae:

$$rh = \sqrt{\text{Rh neg.}} \tag{8}$$
$$Rh' = \sqrt{\text{Rh}' + \text{Rh neg.}} - \sqrt{\text{Rh neg.}} \tag{9}$$
$$Rh'' = \sqrt{\text{Rh}'' + \text{Rh neg.}} - \sqrt{\text{Rh neg.}} \tag{10}$$

If w, u, and v represent the frequencies of genes W, U, and V, respectively,

then

$$w = \sqrt{\text{W}} \tag{11}$$
$$u = \sqrt{\text{W} + \text{U}} - \sqrt{\text{W}} \tag{12}$$
$$v = \sqrt{\text{W} + \text{V}} - \sqrt{\text{W}}, \tag{13}$$

and the frequencies of genes Rh_0, Rh_1, and Rh_2 can now be derived by subtraction, with the aid of EQUATIONS 5 to 7.

Using these formulae, the gene frequencies have been calculated for the available studies on the distribution of the Rh blood types among whites,[14, 15, 16] Negroes, [14, 17] Asiatic Indians,[18] Chinese, [15, 19] Japanese,[20, 21] and Mexican Indians[22] (TABLE 9). In order to determine the statistical significance of the deviation (D) of the sum of the gene frequencies (ΣRh_i) from 100 per cent, the probable error of this deviation was calculated with the aid of Bernstein's formula:

$$\text{P.E.}_{\text{D}} = 0.6745 \sqrt{\frac{uv}{2\text{N}(1 - u)(1 - v)}}, \tag{14}$$

where N equals the number of individuals tested.

It will be seen that, in the majority of investigations, the sum of the gene frequencies falls short of 100 per cent and, in a number of cases, the deviation is significant, in comparison to its probable error. This is explained, in part, by the fact that, in all but the three most recent investigations, the anti-Rh″ reagent was prepared from an anti-Rh_0'' serum which was diluted with saline to eliminate the action of the weak anti-Rh_0 agglutinin which it contained. False positive reactions produced the small amount of anti-Rh_0 agglutinin remaining, and caused some bloods lacking the Rh″ factor to be classified as positive. As a result, an excessive number of bloods were classified as type Rh_1Rh_2. In the case of white individuals, new data are now available which were compiled with the aid of more specific reagents (Wiener and Sonn, series b; and Unger); and the new results fully satisfy the requirements of the six-gene theory. In the case of the Oriental races studied, however, the positive deviations are so high that it is unreasonable to ascribe this entirely to technical imperfections. Here, another complication enters, the nature of which will become apparent when the so-called Hr factor is discussed.

TABLE 9

STATISTICAL TEST OF THE SIX-GENE THEORY

Population	Investigators	Number of individuals tested	Distribution of Rh blood types								Gene frequencies						$D = 100 - \Sigma Rh$	P. $E._D$
			Rh-	Rh_1	Rh_2	Rh_1Rh_2	Rh_0	Rh'	Rh''	Rh'Rh''	*rh*	Rh_1	Rh_2	Rh_0	*Rh'*	*Rh''*		
Whites (N.Y.C.)	Wiener and Sonn	a) 1000	12.9	54.1	12.8	16.4	2.6	0.9	0.3	0	35.9	43.4	13.7	3.5	1.2	0.4	+1.9	±0.5
		b) 667	13.8	56.1	15.6	11.7	1.5	1.2	0.2	0	37.1	44.5	16.4	2.0	1.6	0.2	−1.8	±0.7
	Unger	2434*	14.5	52.5	15.7	13.1	2.4	1.1	0.7	0	38.0	41.5	15.8	3.0	1.4	0.8	−0.5	±0.4
Negroes (N.Y.C.)	Wiener, Belkin, and Sonn	223	8.1	20.2	22.4	5.4	41.2	2.7	0	0	28.4	11.7	14.4	42.1	2.7	0	+0.7	±0.5
Asiatic Indians (Moslems)	Wiener, Sonn, and Belkin	156	7.1	70.5	5.1	12.8	1.9	2.6	0	0	22.6	56.2	6.0	3.4	4.4	0	+3.4	±1.1
Chinese	Wiener, Sonn, and Yi	132	1.5	60.6	3.0	34.1	0.8	0	0	0	12.2	64.2	7.9	2.9	0	0	+12.8	±1.6
Mexican Indians	Wiener, Zepeda, Sonn, and Polivka	98	0	48.0	9.2	41.8	1.0	0	0	0	0	60.0	21.9	10.0	0	0	+8.1	±3.1
Japanese	Waller and Levine	150	1.3	37.4	13.3	47.3	0	0	0	0.7	11.4	50.8	26.8	0	0	0	+11.0	±2.4
Japanese	Miller and Taguchi	180	0.6	51.7	8.3	39.4	0	0	0	0	7.8	64.6	22.1	0	0	0	+5.5	±2.6

* Includes series of 468 bloods, recently reported, in which the tests were made with the improved reagents.

Levine and Javert[23] obtained, from an Rh-positive mother of an erythroblastotic infant, a serum which agglutinated the bloods of 30 per cent of all individuals, independently of the four blood groups. The serum agglutinated all Rh-negative bloods and, for this reason, was designated as anti-Hr, and the corresponding factor as Hr, to indicate its apparent reciprocal relationship to the Rh factor; the serum also agglutinated all Rh-positive bloods which were not agglutinated by anti-Rh′ serum. Race and Taylor,[24] in England, obtained, in a similar case, an antiserum designated by them St (after the first two letters of the patient's surname), which also had the property of agglutinating all Rh-negative bloods. However, the so-called St serum agglutinated the bloods of 80 per cent of all individuals, so these investigators believed their serum to be different from anti-Hr. According to Wiener, Davidsohn, and Potter,[25] the factors Hr and St are identical, and the discrepancy is evidently due to the fact that the serum of Levine and Javert was weak, so that weak positive bloods were erroneously classified as Hr negative in their studies. Accordingly, in the subsequent discussion, the term, Hr, will be used also when discussing the work of the British investigators

The nature of the Hr factor and the mechanism of its hereditary transmission have been clarified by the work of Race and Taylor. According to these investigators, the Hr factor is present in the agglutinogens determined by genes rh, Rh_0, Rh_2, and Rh'', and absent from the agglutinogens determined by the genes Rh_1 and Rh'. Therefore, among the 21 genotypes possible under the six-gene theory, only three genotypes will lack the Hr factor, namely, genotypes Rh_1Rh_1, Rh_1Rh', and $Rh'Rh'$. If we postulate, moreover, that a single dose of the genes determining the Hr factor will cause a weak reaction, while a double dose will determine a strong reaction, then the expected reactions with anti-Hr serum, in relation to the eight Rh blood types, would be as given in TABLE 10. If we combine the bloods which are expected to give strong reactions with anti-Hr serum, these total 32.5 per cent, which is close to the figure of 30 per cent, given by Levine and Javert for their anti-Hr serum. On the other hand, if we now add the bloods expected to give weak reactions with the anti-Hr serum, the total percentage of positive reactions becomes 80 per cent, or the same as that reported by Race and Taylor for their anti-St serum. This supports our contention that St and Hr are the same. It should be mentioned that Dr. Peter Vogel has recently obtained an anti-Hr serum, of exceptional potency, from an Rh-positive mother of an erythroblastotic infant. This serum is of such high titer (250) that, when used in

TABLE 10

THE HR FACTOR IN RELATION TO THE EIGHT RH BLOOD TYPES

Rh blood types	Reactions with Rh antisera			Genotypes	Expected reaction with anti-Hr serum	Frequencies among white individuals (per cent)
	Anti-Rh_0	Anti-Rh′	Anti-Rh″			
Negative	Neg.	Neg.	Neg.	*rhrh*	Strong	14.0
Rh_1 (Rh_0')	Pos.	Pos.	Neg.	Rh_1Rh_1 Rh_1Rh'	Neg.*	20.0
				Rh_1rh Rh_1Rh_0 $Rh'Rh_0$	Weak*	33.5
Rh_2 (Rh_0'')	Pos.	Neg.	Pos.	Rh_2Rh_2 Rh_2Rh'' Rh_2rh Rh_2Rh_0 $Rh''Rh_0$	Strong	15.5
Rh_1Rh_2 ($Rh_0'Rh_0''$)	Pos.	Pos.	Pos.	Rh_1Rh_2 Rh_1Rh'' $Rh'Rh_2$	Weak	13.0
Rh_0	Pos.	Neg.	Neg.	Rh_0Rh_0 Rh_0rh	Strong	2.5
Rh′	Neg.	Pos.	Neg.	$Rh'Rh'$	Neg.	0.02
				$Rh'rh$	Weak	1.0
Rh″	Neg.	Neg.	Pos.	$Rh''Rh''$ $Rh''rh$	Strong	0.5
Rh′Rh″	Neg.	Pos.	Pos.	$Rh'Rh''$	Weak	0.02

* As a rule, Hr-negative type Rh_1 blood is agglutinated more strongly than Hr-positive type Rh_1 blood by anti-Rh′ serum.

routine tests in dilutions as high as 1 to 20, it strongly agglutinates single dose bloods as well as double dose bloods, so that all Hr-positive bloods, both "weak" and "strong," are uniformly strongly agglutinated by this serum.

It is obvious that the discovery of the Hr factor does not make it necessary to change the designations of the six standard Rh genes. These genes can be distinguished by the reactions they determine with anti-Rh_0, anti-Rh′, and anti-Rh″ sera. The anti-Hr serum merely detects another factor in the agglutinogens determined by these genes and is not required to differentiate the agglutinogens from one another. The genes are already clearly identified by the old names, so there is little to gain by further complicating the designations of the Rh genes in order to include their behavior in respect to anti-Hr serum. After all, the purpose of a name is to identify and not to give a full description, though a good name will be based on the salient characteristics of the object named.

The theory of the Hr factor proposed by Race and Taylor, like

Wiener's six-gene theory of the Rh blood types, can be tested by investigations on family material and by the statistical analysis of data on the distribution of the Hr factor in the population. With regard to family material, the following rules of heredity would be expected to hold:

(1) The Hr factor cannot appear in the blood of a child, unless present in the blood of one or both parents.

(2) Hr negative parents cannot have children of types Rh_2, Rh″, Rh_0, and Rh negative, and parents of types Rh_2, Rh″, Rh_0, and Rh negative cannot have Hr negative children.

No exception to either of these rules was encountered in family investigations of Race *et al.*,[26] or in the unpublished studies of the present authors. In fact, we are already applying the Hr tests routinely in medico-legal cases, alongside of the tests for the blood groups, sub-groups, M-N and Rh types.

As shown in TABLE 10, in studies on the distribution of the Hr factor in the general population, Hr negative individuals would be expected to occur practically exclusively in type Rh_1, only rarely in Rh′, and not at all in the remaining six Rh blood types. Race and Taylor found, however, that while, as expected, individuals belonging to types Rh_2, Rh_0, Rh″, and Rh negative were uniformly Hr positive, there were rare individuals of type Rh_1Rh_2 who were Hr negative, contrary to expectation under the theory. This led them to postulate the existence of a seventh allelic gene Rh_y, which, at first, was defined merely in part by its ability to determine a blood property capable of reacting with anti-Rh″ serum, but not with anti-Hr serum. The idea that this new property was genetically determined was confirmed by Race and Taylor,[28] when they encountered a mother and child, both of whose bloods reacted with anti-Rh_0, anti-Rh′, and anti-Rh″ sera, but not with anti-Hr serum. More recently, Murray, Race *et al.*[29] have presented evidence for the existence of an eighth allelic gene, tentatively designated as Rh_z. In conformity with the nomenclature proposed by Wiener, the genes Rh and Rh_z are preferably designated as Rh''' and $Rh_{1\,2}$, respectively. The relationship of these two genes to the six standard genes is shown in TABLE 11.

Since the agglutinogens determined by the genes which confer the ability to react with the agglutinin anti-Rh′ are always Hr negative, while the genes determining the agglutinogens which do not react with the anti-Rh′ serum confer the ability to react with anti-Hr, it is evident that the factors Rh′ and Hr are related to each other genetically,

TABLE 11

REACTIONS WITH ANTI-HR SERUM DETERMINED BY THE SIX STANDARD GENES AND PROBABLE NATURE OF THE SO-CALLED Rh_y AND Rh_z GENES

Genes	Reactions with Rh antisera			Reaction with anti-Hr serum
	Rh′	Rh″	Rh_0	
rh	−	−	−	+
Rh_0	−	−	+	+
Rh′	+	−	−	−
Rh_1	+	−	+	−
Rh″	−	+	−	+
Rh_2	−	+	+	+
Rh′ ″ (Rh_y)	+	+	−	−
Rh_{12} (Rh_z)	+	+	+	−

like M and N. In fact, Wiener, Davidsohn, and Potter have shown that if tests are made only with anti-Rh′ and anti-Hr sera, three types of blood can be distinguished which are analogous, serologically and genetically, to the three M-N types.

Recently, we have been able to confirm the reports of Race *et al.* that there exist one or two genes (*Rh′ ″* and/or $Rh_{1\,2}$) determining agglutinogens reacting with both Rh antisera, anti-Rh′ and anti-Rh.″ This gene (or genes) is evidently rare among white individuals, because, among 1200 persons, there were only two whose blood gave reactions corresponding to type Rh_1Rh_2 and who were, nevertheless, Hr negative.[30] In one case, the child of the individual in question also belonged to type Rh_1Rh_2 and was Hr negative, just as in the case reported by Race and Taylor, supporting the view that a special gene was involved. In view of the rarity of the genes *Rh′ ″* and/or $Rh_{1\,2}$ among white individuals, it is not surprising that, in the statistical test of the six-gene theory of the Rh blood types, the sum of the gene frequencies did not deviate significantly from 100 per cent. In recent investigations of Mexican Indians,[22] however, evidence was obtained that, in this race, the genes *Rh′ ″* and/or $Rh_{1\,2}$ are far more common than among white individuals. Thus, among 98 Mexican Indians, there were as many as three individuals of type Rh_1Rh_2 and Hr negative. In view of the evidence that American Indians are of Mongolian derivation, it may be presumed that the genes *Rh′ ″* and/or $Rh_{1\,2}$ have a similar incidence among Chinese and Japanese. This would explain why, for these races, the calculated sum of the gene frequencies under the six-gene theory falls short of 100 per cent.

Obviously, bloods of genotypes *Rh′ ″*Rh_0, $Rh_{1\,2}Rh_0$, $Rh_{1\,2}rh$, if they

exist at all, should give the same reactions with antisera Rh_0, Rh′, Rh″, and Hr as typical bloods of type Rh_1Rh_2. Accordingly, the existence of genes *Rh′″* and/or $Rh_{1\,2}$ leaves open the possibility of exceptions to the second rule of heredity, under the six-gene theory. That such exceptions have not been encountered, to date, may be explained by the rarity of these genes. At any rate, it is still justifiable to conclude, in the case of exclusions under the second rule, that paternity (or maternity) is highly improbable, even though it is not necessarily impossible. The situation is somewhat similar to that which exists with regard to Bernstein's three-gene theory of heredity of the four blood groups, O, A, B, and AB. One or two exceptions to the second rule of heredity have been encountered, apparently not attributable to technical errors or illegitimacy. As Wiener[31] has pointed out, such exceptions can be explained by postulating the existence of a fourth gene which produces the same effect as genes *A* and *B* together; or what amounts to the same thing, by postulating heredity by a system of four completely linked genes, (*ab*), (*Ab*), (*aB*), and (*AB*).

Fisher[32] has postulated the existence of two additional antisera and corresponding factors, which are related to Rh″ and Rh_0, respectively, in the same way that Hr is related to Rh′. On the other hand, Wiener suggests that Hr holds an unique position in the scheme of the Rh factors, analogous to the position of factor O in the scheme of the A-B-O agglutinogens (cf. Wiener and Karowe[14]), a view which would preclude the existence of the two additional factors postulated by Fisher. To date, no convincing evidence has been reported of the existence of antisera or factors corresponding to Fisher's hypothetical factors. In any event, Fisher's suggestion to change the nomenclature completely, in order to conform with his hypothesis, is unwarranted, because the present nomenclature can easily be extended to include the factors he has postulated (TABLE 12). Another objection is that Fisher's designations of the genes and corresponding agglutinogens are cumbersome and deviate considerably from the usual practice of geneticists. The main objection is that three new letters and symbols are introduced, which have no relation to the symbol Rh.

In conclusion, it should be mentioned that there is evidence for the existence of genes in the Rh series, in addition to the eight genes discussed in the present presentation. These genes are characterized by the fact that the agglutinogens which they determine give weak or intermediate reactions with one or more of the Rh antisera.[35] As shown by Wiener, these genes are rare and occur most frequently among Negroes.

TABLE 12

HOW AUTHORS' DESIGNATIONS CAN BE EXTENDED TO INCLUDE FISHER'S HYPOTHETICAL FACTORS

Genes: Wiener	Genes: Fisher	Designations of antisera (Fisher / Wiener): Γ Rh'	H Rh''	Δ Rh_0	γ Hr' Hr	η (Hr'')	δ (Hr_0)
rh	*cde*	−	−	−	+	(+)	(+)
Rh'	*Cde*	+	−	−	−	(+)	(+)
Rh''	*cdE*	−	+	−	+	(−)	(+)
Rh_0	*cDe*	−	−	+	+	(+)	(−)
Rh_1	*CDe*	+	−	+	−	(+)	(−)
Rh_2	*cDE*	−	+	+	+	(−)	(−)
Rh' ''	*CdE*	+	+	−	−	(−)	(+)
$Rh_{1\,2}$	*CDE*	+	+	+	−	(−)	(−)

SUMMARY

The evolution of knowledge concerning the Rh blood types has been described. To date, there is evidence for the existence of three Rh factors and one so-called Hr factor. The proposal to designate the Rh factors Rh′, Rh″, and Rh_0, respectively, and the special designation, Hr, for the Hr factor, are based on the serologic and genetic properties of these factors. The designations proposed for the types were devised from the point of view of simplicity, as well as conformity with the serologic and genetic facts.

The available family and statistical data were examined, with regard to the theory of six standard allelic Rh genes, and found to support the theory in the main. The rare deviations from the theoretical expectations are explained by postulating the existence of two additional allelic genes, *Rh′ ″* and $Rh_{1\,2}$, corresponding to the genes *Rhy* and *Rhz* of Race *et al.* These genes appear to be rare among white individuals, but occur more frequently among Mexican Indians and presumably, therefore, also among Chinese and Japanese. The problems raised by the existence of these genes, in relation to the genetic theory and the problem of nomenclature, were discussed.

The alternative methods of designation proposed by Murray and Fisher were discussed. Since these designations offer no special advan-

tage and are more complicated than those proposed by Wiener, there seems to be no good reason for making any change, at least for the time being. As is pointed out, the present designations are quite flexible and can readily be extended to include any new Rh or Hr factors which may be found in the future.

BIBLIOGRAPHY

1. **Landsteiner, K., & A. S. Wiener**
   1940. Proc. Soc. Exp. Biol. & Med. **43:** 223.
2. **Wiener, A. S.**
   1943. Blood Groups and Transfusion: 254. Third edition, C. C. Thomas. Springfield, Ill.
3. **Murray, J.**
   1944. Nature **154:** 701.
4. **Landsteiner, K., & A. S. Wiener**
   1941. J. Exp. Med. **74:** 309.
5. **Wiener, A. S.**
   1944. Science **99:** 532.
6. **Wiener, A. S.**
   1945. J. A. M. A. **127:** 294.
7. **Wiener, A. S.**
   1941. Arch. Path. **32:** 227.
8. **Wiener, A. S., & K. Landsteiner**
   1943. Proc. Soc. Exp. Biol. & Med. **53:** 167.
9. **Wiener, A. S.**
   1943. Science **98:** 112.
10. **Wiener, A. S., & E. B. Sonn**
    1943. J. Immunol. **47:** 461.
11. **Race, R. R., G. L. Taylor, K. E. Boorman, & B. S. Dodd**
    1943. Nature **152:** 563.
12. **Wiener, A. S.**
    1943. Proc. Soc. Exp. Biol. & Med. **54:** 316.
13. **Wiener, A. S., & E. B. Sonn**
    1945. J. Lab. & Clin. Med. **30:** 395.
14. **Wiener, A. S., E. B. Sonn, & R. B. Belkin**
    1943. J. Exp. Med. **79:** 235.
15. **Wiener, A. S., E. B. Sonn, & R. B. Belkin**
    1943. Proc. Soc. Exp. Biol. & Med. **54:** 238.
16. **Wiener, A. S., L. J. Unger, & E. B. Sonn**
    1945. Proc. Soc. Exp. Biol. & Med. **58:** 89.
17. **Wiener, A. S., R. B. Belkin, & E. B. Sonn**
    1944. Am. J. Phys. Anthrop. **2** (N. S.): 187.
18. **Wiener, A. S., E. B. Sonn, & R. B. Belkin**
    1945. J. Immunol. **50:** 341–348.
19. **Wiener, A. S., E. B. Sonn, & C. L. Yi**
    1944. Am. J. Phys. Anthrop. **2** (N. S.): 267.
20. **Waller, R. K., & P. Levine**
    1944. Science **100:** 453.
21. **Miller, E. B., & T. Taguchi**
    1945. J. Immunol. **51:** 227–231.
22. **Wiener, A. S., J. P. Zepeda, E. B. Sonn, & H. R. Polivka**
    1945. J. Exp. Med. **81:** 559.

23. Levine, P.
1943. J. Ped. **23**: 656.

24. Race, R. R., & G. L. Taylor
1943. Nature **152**: 300.

25. Wiener, A. S., I. Davidsohn, & E. L. Potter
1945. J. Exp. Med. **81**: 63.

26. Race, R. R., G. L. Taylor, E. W. Ikin, & A. M. Prior
1944. Ann. Eugen. **12**: 206.

27. Race, R. R., G. L. Taylor, D. F. Cappell, & M. N. McFarlane
1944. Nature **153**: 52.

28. Race, R. R., & G. L. Taylor
1944. Nature **153**: 560.

29. Murray, J., R. R. Race, & G. L. Taylor
1945. Nature **155**: 112.

30. Wiener, A. S., E. B. Sonn, & L. J. Unger
Unpublished observations.

31. Wiener, A. S., M. Lederer, & S. H. Polayes
1930. J. Immunol. **17**: 218.

32. Wiener, A. S.
1943. Blood Groups and Transfusion: 185. Third edition. C. C. Thomas. Springfield, Ill.

33. Fisher, R.
1944. Nature **153**: 771. (Cited by **R. R. Race.**)

34. Wiener, A. S., & H. Karowe
1944. J. Immunol. **49**: 51.

35. Wiener, A. S.
1944. Science **100**: 595.

DISCUSSION OF THE PAPER

Dr. R. R. Race (*Cambridge, England*):*

First, there has been far too much reiteration on this subject, by both British and American workers. Our entire contribution to this subject has been contained in about half a dozen letters to *Nature*, but it seems that these have not been properly read, and this is our own fault for having indulged in reviews, etc.

The American writing on the genetic aspect of Rh is much more cautious than that of the Fisher school, here. We have gone further than Wiener, and much further than Levine has ventured. Wiener and Levine probably think we have gone so far that we are lost, and, of course, they may be right. This is roughly the position here:

Eighteen months ago, we (Race, Taylor, Cappell, McFarlane, Boorman, and Dodd) reached this stage (TABLE 1), in agreement with, but independently of, Wiener. Wiener's work lacked the very great help of the St serum. We brought our notation into line with Wiener's.

TABLE 1

Genes	*Anti-Sera*
Rh_1 becoming Rh_0 Rh_2 becoming Rh'' rh becoming Rh'	KJ became anti-Rh″

* By correspondence. (Some slight changes in nomenclature were made by Dr. Boyd.)

TABLE 2

	Wiener Antisera							Race *et al* Antisera			
Genes	Rh	(St)	Rh_2	Rh′	Rh_1	Rh″	Genes	Rh	St	Rh″	Rh′
Rh_1	+	0	−	+	+	+	Rh_1	+	−	−	+
Rh_2	+	0	+	−	+	+	Rh_2	+	+	+	−
Rh	+	0	−	−	+	+	Rh_0	+	+	−	−
Rh′	−	0	−	+	+	−	*Rh′*	−	−	−	+
Rh″	−	0	+	−	−	+	*Rh″*	−	+	+	−
rh	−	0	−	−	−	−	*rh*	?	−	+	?

TABLE 3*

Name of serum	Anti-Rh′	Hr	Anti-Rh_0	Anti-Rh″	Not yet found	
Antibody present.	Γ	γ	Δ	H	δ	η
Genes						
Rh_z CDE	(+)	(−)	(+)	(+)		
Rh_1 CDe	+	−	+	−		
Rh_y CdE	(+)	−	(+)	+		
Rh Cde	+	−	−	−		
Rh_2 cDE	−	+	+	+		
Rh_0 cDe	−	+	+	−		
Rh″ cdE	−	+	−	+		
rh cde	−	+	−	−		

* Those reactions not yet determined serologically are given in parentheses

In January, 1944, Fisher examined TABLES 1 and 2 and noticed that, when a gene was positive with St, it was negative with anti-Rh′. He supposed that the antigens disclosed by these two sera were due to alleles, and these alleles he called c and C. To the antibodies, he gave the corresponding Greek letters, St becoming γ, and anti-Rh_1, Γ. He considered that the antigen recognized by anti-Rh_0, which he called D, had an allele d, the antiserum for which was yet to be found. The antigen recognized by anti-Rh_2, he called E, and this serum he called H. The η antiserum, for the hypothetical allele e, was also to be found.

In other words, there were three tightly linked *loci* on the chromosome responsible for Rh, making eight possible combinations in a chromosome (TABLE 4).

TABLE 4

Antisera	*Genes**	*Antisera*
(Anti-Rh′)	C or c	γ (St) (Hr)
(Anti-Rh_0)	D or d	δ } then to be found
(Anti-Rh″)	E or e	η

*D = Rh_0
C = *Rh′*
E = *Rh″*

This scheme I had the privilege of publishing in a letter to *Nature*, in June, 1944. Later, a segregation in the offspring of the first "Rh_y" person I had found showed that the reactions (previously marked by the other genes present) with Δ

and Γ were those booked by Fisher for Rh_z. So, unfortunately, we had to change the name from Rh_y to Rh_z. Rh_y remains undiscovered, unless Wiener has found it amongst the Mexican Indians. Fisher's scheme predicted that both these genes must be Γ +, and the one found turned out to be so.

The rest of the story is told in two letters to *Nature* which appeared Saturday, May 5th.

Briefly, Mourant has found η, and the anti-Hr, described by Waller and Levine (Science. 100: 453. 1944), has the reactions predicted by Fisher for δ.*

We have published the results on 56 families tested with the four anti-Rh sera, and those on 44 more have been in the press some time (Ann. Eugen.).

Quite a number of workers in England are using St serum, together with the other three, in the routine examination of the blood, at any rate of the fathers of erythroblastotic infants.

Wiener confirmed in full our work with the St serum, but he still speaks always of the eight Rh types, ignoring the distinctions made by St. It is difficult to see how Wiener's intermediates fit into Fisher's scheme. I think, myself, the scheme is correct, as far as it goes, but that the situation is probably still more complicated. However, it is a beautiful piece of prophetic biology.

About nomenclature, my opinion, for what it is worth, is that, for clinical purposes, Rh positive and Rh negative seem to be very satisfactory, qualified, where necessary, with the genotype, *e.g.,* Rh_1Rh_1, $Rh'rh$. But, for more scientific purposes, assuming that Fisher's scheme becomes accepted, I think the allelomorphisms should be indicated by the names. On Fisher's notation, Rh_1rh, for example, is CDe/cde.

I hope all this does not look as if I had lost sight of the great debt we all owe to Wiener and Levine. The English work has only been a small extension, I hope a correct one, of their fundamental discoveries.

Dr. Polayes:

The pathologist happens to be in a practical position midway between the more elevated scientist, represented by this group, and the general practitioner. Being in that position, I can tell you definitely that any attempt to select a terminology satisfactory to both the geneticist and the practitioner (for whom Dr. Levine has justifiably made his plea) is futile. I am convinced, nevertheless, that simplification of the present terminology is most essential for practicability. Might I suggest, in that direction, since the terminology which states the percentage of positives for each of the Rh factors has already proved to be practicable, because of its descriptive character, that, for the time being, at least, this terminology be accepted for the medical profession at large, and that the special students in this field adopt a terminology more suitable for themselves. Thus, most physicians will readily understand Rh (85%) to mean that this particular antigen is positive in 85 per cent of a mixed white population; Rh (70%), in 70 per cent; Rh (87%) in 87 per cent, etc. Simplification, at this stage of development of the subject, is most essential for practicability.

Dr. Boyd:

It seems possible that it will be necessary to face the facts, in the Rh situation, just as many medical practitioners have reluctantly admitted the existence of at least two kinds of A antigen, in A and AB cells. In this case, it has not yet proved necessary, in general, to do much about the distinction. In the case of Rh, where a characteristic disease entity is produced, it may be necessary to use a terminology more detailed than many medical men would like. But we are not in a position to dictate to nature how complicated we will allow scientific fact to be.

Dr. Philip Levine:

I should like to ask Dr. Wiener several questions regarding his theory on the heredity of the Rh factor.

1. Are the results with Dr. Wiener's anti-Rh″ agglutinin entirely reliable?

* This apparent agreement has since been claimed by Levine to be due to a typographical error in his own paper.

Waller and I found that dilution of Dr. Wiener's serum did not entirely remove the action of the concomitant anti-Rh_0 agglutinin (Science. 100: 453. 1944). How big is the error, due to residual action of anti-Rh_0, in his heredity studies? Would Dr. Wiener recommend the use of anti-Rh″ agglutinin in this particular serum, which is a mixture of two agglutinins, in cases of disputed paternity?

2. Are Dr. Wiener's findings with the anti-Hr serum entirely reliable? He states, in a paper with Davidsohn and Potter, that all activity disappeared in the course of the study. I assume that it never was very active, and, in a specimen sent to me by Dr. Davidsohn, it was entirely inactive. After my experience with anti-M and anti-N sera, it is difficult for me to appreciate how one can carry out heredity studies with such an inactive agglutinin and yet present figures indicative of sharp differentiation of bloods possessing and lacking the factor. Did Dr. Wiener experience any difficulty in reading his tests with this weak anti-Hr serum?

3. In a recent paper (Science. 100: 595. 1944) Dr. Wiener writes that "anti-Hr sera have a place in the scheme of Rh blood types similar to that of the anti-O sera in the blood group scheme." In support of his view, he notes that ". . . anti-O sera, like anti-Hr sera, are unusually of low potency." Does Dr. Wiener agree with me, now, that anti-Hr sera are weak because of statistical and not genetic necessity? Only about 2% of all mothers of erythroblastotic infants are Hr negative and can produce anti-Hr agglutinins, as compared to 92% Rh— women who can produce anti-Rh agglutinins. It is, therefore, obvious that the possibility of finding potent anti-Hr sera is 46 times less than of finding potent anti-Rh sera. All told, I studied three potent anti-Hr sera and one weak one Of the four sera, the first one studied gave the weakest reactions. An anti-Hr serum, supplied to me recently by Dr. Vogel, had a titer of 1:512. In the light of these findings, does Dr. Wiener still hold to the view that the relationship of the Hr factor to the Rh is analogous to that of the subgroup of A to the four groups?

4. Finally, any terminology (and there are already several) is, in a sense, arbitrary. So far as the clinician is concerned, he need not bother with any other terminology but that defined by the anti-Rh_0 serum. At the same time, one can recommend to him the simple genetic theory, based on a single agglutinin which resolves 92% of the cases. In a recent paper, to appear in the June issue of the Am. J. Obstetrics and Gynecology, I make this suggestion: For the exceptional 8% Rh+ mother, the obstetrican should be in touch with a specialist in the field.

Does Dr. Wiener agree with me that it is essential to avoid confusing the clinician with complicated schemes, based on rare, or impure, or completely hypothetical antibodies? No genetic theory can be considered final, until potent reagents of all varieties are available and the role of the Hr factor in the genetic scheme is defined. The final terminology will have to be decided upon by an international committee of geneticists and serologists.

Dr. A. S. Wiener:

Dr. Levine's questions may be answered as follows:

1. The answer to his first question can be found in my paper mentioned in his third question (Science. 100: 595. 1944). "It should be mentioned that the anti-Rh′ and anti-Rh″ sera used in these studies were really anti-Rh_0′ and anti-Rh_0″ sera with weak anti-Rh_0 agglutinins whose action was eliminated by simple dilution. Recent studies reveal that some bloods containing Rh_0 factor react distinctly with traces of anti-Rh_0 agglutinin insufficient to clump other bloods containing Rh_0 factor. In this way, reactions could occur due to residual anti-Rh_0 agglutinins in the serum, which might erroneously be attributed to the anti-Rh′ and anti-Rh″ agglutinin. This pitfall can now be eliminated with the aid of potent anti-Rh_0 blocking serum, because if such a serum is mixed with diluted anti-Rh_0′ or anti-Rh_0″ serum, the action of the anti-Rh_0 agglutinin will be completely eliminated." In a subsequent study, using these improved reagents (Proc. Soc. Exp. Biol. & Med. 58: 89. 1945), it was established that the earlier studies, with the less specific reagents, contained only a 3 per cent error in the tests with the anti-Rh′ serum, and a negligible error, if any, in the tests with the anti-Rh″

serum. Naturally, this very small error did not in any way affect our conclusions as to the hereditary mechanism of the eight Rh blood types. With regard to the use of the anti-Rh″ agglutinin in the serum, to which Dr. Levine refers for cases of disputed paternity, this is perfectly reliable, now that anti-Rh_0 blocking serum is available, provided that the experimenter has sufficient experience with these delicate tests.

2. The results of our Hr tests speak for themselves, as can be seen from a careful study of our paper (J. Exp. Med. 81: 63. 1945). The serum originally had a titer of 10 to 20, so that, when we used it, our results were clean-cut. Furthermore, in subsequent tests, with a very potent anti-Hr serum (titer 250) obtained from Dr. Peter Vogel, very similar results have been obtained.

3. As Dr. Levine points out, in my paper (Science, 100: 595. 1944), I state that "anti-O sera, like anti-Hr sera, are *usually* of low titer." Dr. Levine apparently failed to notice that I did not say "invariably," because sera, like the Hr serum of Dr. Vogel, mentioned above, make it necessary to use the word "usually." With regard to my idea that ". . . anti-Hr sera have a place in the scheme of the Rh blood types similar to that of the anti-O sera in the blood group scheme" (and, I might add, similar to that of anti-N sera in the M-N scheme), this concept has proved very useful to me in explaining the reactions of Hr antisera. In the development of such a complex subject as the Rh factors, it is inevitable that some of the earlier hypotheses should later prove to be somewhat inaccurate Even an inaccurate hypothesis is of value, however, if it makes progress possible and helps one to understand the subject As soon as a better explanation of the reactions of Hr antisera is offered, I shall be glad to abandon my hypothesis, because then it will have outlived its usefulness. Until then, it may prove helpful to those desiring to grasp the fundamentals of this complicated subject (cf J. Immunol. 49: 51. 1944).

4. It is true that any nomenclature is, in a sense, arbitrary. However, a logical nomenclature for the Rh blood types must take into account the special positions of the anti-Rh_0 and anti-Hr agglutinins, and the fact that the anti-Rh′ and anti-Rh″ agglutinins are related to one another, like anti-A and anti-B. The nomenclature I have devised takes these facts into account.

I agree that, for most clinical problems, sufficient information will be gleaned from the mere classification of bloods as Rh positive or Rh negative, with the aid of standard anti-Rh_0 serum. However, when an expert on the subject is called into consultation on a specific problem, as I often am, the clinician will expect him to give more refined information than can be obtained from the ordinary clinical pathologist. A complete report should include data as to the Rh blood types, the Hr factor, tests for sensitization by the tube agglutination test, blocking test, slide test, and tube conglutination test, when these offer more specific information as to the diagnosis, prognosis and treatment of the patient.

The nomenclature I have proposed has proved workable and has already been widely adopted. I agree that the final terminology will have to be agreed upon by an international committee of geneticists and serologists, but this should not differ in its main outlines from that now in use.

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES
VOLUME XLVI, ART 10. PAGES 993–1122
JANUARY 15, 1947

BRAIN AND BODY WEIGHT IN MAN: THEIR ANTECEDENTS IN GROWTH AND EVOLUTION *

A Study in Dynamic Somatometry

BY

EARL W. COUNT

*Professor of Anthropology, Hamilton College, Clinton, N. Y.;
Research Associate, The Viking Fund, Inc., New York, N. Y.*

* Publication made possible through a grant from the Viking Fund and the Annals Revolving Fund.

PREFACE

During the past half-century, the growth of the human brain and the changing relations of brain weight to body weight, both in man and other animals, have been investigated by numerous authors, notably Dubois, Lapicque, Donaldson, Brummelkamp, von Bonin, and many others. But, while these studies have established many facts, the underlying principles have not yet been fully revealed. My own studies have approached the classic problem of the relations of brain weight to body weight from the direction of physical anthropology, with the results set forth in the present paper.

For the most part, the data were brought together in the American Museum of Natural History, where I am under obligation to many, and very especially to Dr. W. K. Gregory. Dr. H. B. Latimer of the University of Kansas did me the extraordinary favor of copying his extensive and unique series of measurements on body and brain weights of fetal and adult cats. I wish also to thank Dr. R. W. Miner for his editorial work in preparing the manuscript for printing.

Acknowledgment is due Miss Lillian Guglielmi, Secretary of the Department of Anatomy, New York Medical College, for typing the manuscript.

I am especially grateful to President R. C. Hunt and Dr. Paul Fejos, Director of Research, both of the Viking Fund, Inc., for their sustained interest in this work, and jointly to the Viking Fund and The New York Academy of Sciences for defraying the expenses of publication

CONTENTS

SECTION I. INTRODUCTION

Our general subject is that of the relationship between brain weight and body weight in man. More accurately, we are inquiring whether the relationship in man is at all peculiar and aberrant, with respect to all his relatives in the mammalian class, or whether he fits into a common scheme.

This question breaks down into three problems:

(1) How does brain weight grow with respect to body weight, during the entire period of development of the individual? This is a question of developmental anatomy, from embryo to adult.

(2) How has brain weight changed with respect to body weight during man's evolution, starting at a time when his ancestors were in no wise separable from those of other modern primates? This is a question of paleontology.

(3) What regularity is there, if any, in the pattern obtaining today among mammals; *i.e.*, can a scheme be found which places man, with respect to the brain weight/body weight ratio, in a position understandable in the light of the positions of all other mammals? This is a question of comparative anatomy.

It is no longer disputed that these questions are mutually related. The present study seeks the way to a law, expressible in number, that will describe the relationship.

A hundred years ago, only the first and third questions would have been asked. Today, both the first and third are subordinated to the second, for reasons well known to every biologist. However, unfortunately, we still lack the kind of data that could turn this study into an attempt to answer the second question. Therefore, we shall pursue questions one and three, with the faith that any answer to them will furnish ground work for an eventual answer to question two.

A. ORIENTATION

The objectives of the present study are: (1) To formulate the growth behavior between brain and body weights in human ontogeny; (2) to find any traits in common with other Primates and Mammalia; (3) to formulate analogous behaviors in comparative anatomy; and (4) to

search for any connections between ontogeny and comparative anatomy.

1. The salient features of the ontogenetic aspects: Where $Y =$ log brain weight, $X =$ log body weight, $A =$ log a; both fetal and postinfantile growth behavior fit $Y = A + bX$, as a first approximation; b being much greater, and A being much smaller, in the fetal case than in the other. A transitional curvature connects the two. The sexes are formulated separately.

2. Man is compared with apes, monkeys, cattle, cats, and rats. The factors determining ultimate weight and proportion of brain to body are: (a) Its relative growth velocity before the fetus stage; (b) velocity during the subsequent period, for which we have data; (c) the amount of growth energy expended by the total body per morphological stage achieved; (d) the length and (e) degree of curvature to the line of the transitional period; (f) the length and (g) velocity of the line of postinfantile growth.

Man and ape may together be superior to the monkeys in point (a). They may be slightly superior in point (b), but this is more distinctive of the primates together with respect to the other mammals studied. Man stands highest in point (c), the apes are next, the monkeys third. The apes are highest in point (f). In point (g), the lowest velocity is that of the apes; then comes man; then the monkeys. Apes and man together are distinctive in point (g), while the monkeys resemble other mammals.

3. The salient features of the comparative anatomical aspects, assuming the Insectivora to represent the "mother order" of the extant mammals: (a) The Dubois formula $E = kP^r$ for relating brain weight to body weight is considered invalid. (b) The von Bonin formula (a rectilinear regression as a statistical mean for the whole mammalian class) is considered to fail also, but for a different set of reasons. (c) (1) Treating the mammals separately by orders, each order apparently describes, in terms of log brain weight Y and log body weight X, a parabola $Y = A + bX - cX^2$. (This is termed the "cephalization exponent.") So does the surviving mass of the class Mammalia. (2) The constants b and c are, in every instance, characteristic; but when all instances are assembled, there emerges the relationship $c = f(b)$, where $f(b)$ is rectilinear and negative, and the correlation $r_{bc} = -.912$. (3) The extant reptilian behavior is $Y = A + bX + cX^2$; and b, c agree with the mammalia in $c = f(b)$. (4) The intersection of the reptilian exponent with the general mammalian can be used as the

origin of a new pair of axes, to which all exponents can be referred, thus making $A = 0$. Then, the mammalian exponents can be treated as *transformations* out of the reptilian.

4. The salient features of the ontogenetic and comparative anatomical aspects, when juxtaposed: (a) For a major period of fetal growth, the ontogenetic line approximately parallels the comparative-anatomic (in the orders studied), although the evaluation of the former, rectilineårly, and the latter, parabolically, at the outset precludes complete parallelism (the parabolas in this stretch are gradual); (b) the fetal ontogenetic line tops the comparative-anatomic, so that, per total body weight, the fetus has devoted a larger percentage to brain weight than a putative phyletic ancestor of like body size, indicating the prefetal period as largely responsible for the precocity of brain weight; (c) the steeper the fetal ontogenetic and the comparative-anatomic lines, the greater this precocity; (d) man is not aberrant, but a quite orthodox projection to a larger absolute size of a phylogenetic formula applying equally well to the monkeys, both New and Old World.

A principle underlying all mammalian cephalization is: Increase in absolute body size is a requisite for higher cephalization.

B. STANDPOINT AND METHOD

When an anthropologist measures the body of man and compares its proportions in terms of indices, he commits himself in two ways: He pledges himself to the operations of numbers; or, he assumes an attitude towards the object of his scrutiny.

The pledge is often hard to keep. The ways of the mathematical country are very unexpected, and the traveler finds himself unequipped for rough going. Too frequently, he stops far short of where he might go, either because he finds himself unready or, alas, because he may not be aware of what good things lie beyond.

By far the bulk of anthropometric studies are content with measurements of a particular *state* obtaining for an individual or a group. They may, however, go beyond: they may register and collect, at suitable, successive intervals, the stages of growth for an individual or for a population. By interpolation, a continuum can be filled out. It can be graphed as a curve. Occasionally, the student describes the curve in terms of a formula (empirically). But commonly, the practice is to seek the average of a ratio—the average of an index—for a population at but one (usually definitive) stage of its somatic proportions. Now, ideally speaking, an index can be plotted on a Cartesian

grid in terms of its component measurements, y and x. It is a point. Its only property is position, a perfectly static matter. But how did it get there? Clearly, by tracing a locus. That took time. Explicitly or implicitly, time is there as a parameter. If, now, we can formulate a locus, we have passed from a static to a dynamic method. This is right and proper, for anatomy itself is a process, not just a state. Its behavior is an event. An index, therefore, should be a locus, not a point; a formula in the variables (y, x), or (writing time as t) in (y, t) or (x, t).

To put it more technically: Let an organism grow from embryo to adult. Let a chosen set of stages be represented as:

$$A_1 - B_1 - C_1 - \ . \ . \ . \ Z_1,$$

and plotted.

Let Z be the adult. Let it be understood, however, that there is a continuum between the stages and that the stages are artifices of choice.

Let this organism be an ancestor to a later form with subscript having its homologous set of ontogenetic stages:

$$A_2 - B_2 - C_2 - \ . \ . \ . \ Z_2.$$

Let there be n such phyletic stages, so that the modern form is:

$$A_n - B_n - C_n - \ . \ . \ . \ Z_n.$$

Let us plot all Z's. We find that they conform to a pattern, to a formulable line. Yet Z_1 did not give rise to Z_n. There is no continuum between Z's, as there has been in the other case. It is the germ plasm of the organism with subscript $_1$ that eventuated in the organism with subscript $_n$ by a phylogenetic process. If, instead of Z's, our fossil records housed the stages A, C, or H of a phyletic line, we could make just as legitimate comparisons. However, the formula for the line connecting all C's would not have the same constants as that connecting all H's.

We could, if we had all stages, construct a master-chart in which we connected every A-B-C- . . . Z series with lines, and also every letter-plus-subscript series A_1-A_2-A_3- . . . A_n, etc., with another set of lines. The first would be ontogenetic; the second, phylogenetic. The one set of lines would cut across the other. With such a gridlike chart, we could visualize the gradual *distortion* of an organism during its ontogeny, or of a phyletic line during its evolution.

It is safe to say that this will never be possible. Nevertheless, I hope the belief is not too sanguine that, out of today's qualitative anatomy, with its facets paleontologic, developmental, or comparative in any other form, there will gradually emerge a quantitative anatomy of analogous syncretism, in which *processes* and *form changes* will be pa-

tiently measured and formulated, until reasonable reconstructive extrapolations or interpolations can be made that will further the accuracy of our evolutionary descriptions. It is hoped that the present study will do its bit in that cause.

C. PRINCIPLES AND TERMINOLOGY

1. The plotted points in a graphed field (the "scatter diagram") will be called a *constellation*.

2. If a constellation shows some general organization, a tendency or trend, as though streaming in some direction, we shall speak of an *orientation*, and the line of means of that orientation as the *axis*. The more pronounced the tendency to orient, the higher the correlation of the variables. We may say, then, that the line of mean or mass tendency indicates the *dependence* of one variable y upon the other x, while the amount of dispersal or scatter gives the degree of *independence* between the two.

3. We shall examine constellations of genera, families, orders, and the class Mammalia. The more comprehensive the constellation, the more pronounced the tendency for y to depend upon x the more clearly defined the orientation. The reason is not hard to see The total mammalian constellation is comparatively long and narrow, a species constellation will usually be short and relatively broad. Biologically interpreted, in a species or genus the range of brain size is larger, relative to the range in body size, than in an order or the class

4. When we come to treat of comparative anatomy, we shall examine, first, the system developed by Dubois, Lapicque, Brummelkamp, *et al.*, and accepted and further exploited by others. The system claims for all vertebrates, no matter what their proportion of brain to body weight, a certain constant feature termed the "relational exponent," and a certain feature that varies according to the degree of "cephalization"; *i.e.*, in every vertebrate, the brain weight, y, is related to body weight as $y = ax^b$, where $b = .56$ and is the relational exponent; while a varies according to the "cephalization" of the animal, and is termed the "coefficient of cephalization."

We shall present the counterposed system of von Bonin, and shall attempt to analyze both. In the end, we shall propose another system, one having a "cephalization exponent," and in which the brain and body weights shall be related as $y = ax^{b + c \log x}$. We shall then explore the consequences of its adoption.

The procedure really is simple. We start with man Frankly, he

remains, throughout, the center of interest—human egoism carried over into science. But, in part, we redeem the reproach by widening the angle of vision to include the rest of the primates. Eventually, we set the scope to sweep in other orders In the phylogenetic section, we can, at last, see the class Mammalia; anyway, the remnant that still survives. And beyond, we catch a glimpse of the reptilian background We start with ontogeny (Section II), then switch to comparative anatomy (Section III), and finally essay a comparison of the two (Section IV).

SECTION II. ONTOGENY

A. THE CHARACTER OF THE DATA

Using brain weight at all is, at best, an uncertain procedure. The data come from a wide range of sources, which means strong individualism in securing them. The brains may have been weighed fresh, or after various periods of preservation in all sorts of solutions. They may even have been severed at different levels. This factor can be weighty when the brains are very small. Brains will have been weighed with or without meninges, the cisternae more or less drained. Even under ideal circumstances, it still would be uncertain just how much living substance a brain weight would represent. Aside from the varying amounts of blood in the small vessels, the age and kind of animals means varying amounts of water, although this criticism would touch body weight, as well. Also, there are individualisms in the techniques of weighing which may be very important, especially when small brains are being weighed.

If brain weight is an uncertain datum, so is body weight Some animals are well fed, others emaciated. Monkeys, for one group, are notorious gluttons when they have a chance. Stomach and intestinal content sometimes are an astounding percentage of gross body weight

In the case of man, the data come from the least representative members of the species the minority who have died prematurely. Disease reduces body weight, but brain weight much less profoundly. This is very important in studies limited to man The importance is enhanced when the human are compared to animal data having another set of antecedents.

The kind of data limits their quantity and extent. (If our problem were popular, it would lead to an appalling slaughter of the "lesser creation") Sometimes, in fact, the series for an animal is limited to one specimen. There are no real brain weights at all for *Sinanthropus* or *Homo rhodesiensis*, nor any body weights, either. If we were to substitute cranial capacity, the horizon would immediately recede a distance.

Cranial capacity, to be sure, may be used to estimate the weight of the brain it once contained. But aside from the technical difficulties

preventing, in their turn, a standardized measure, one cannot be sure of the modulus by which to turn the cubic centimeters into grams. In closely related forms, like *Pithecanthropus* and *Homo*, a collateral study of living primates may permit a fairly safe judgment in the matter. However, when the range of comparison is widened taxonomically, uncertainty increases.

Also, if one is studying the *Palaeanthropi*, etc., one must estimate body weight as well.

All of these bafflements have resulted in some students advocating the use of cranial capacity, nevertheless, as the lesser of two evils. The students of fossils go further, since they have, at best, only bones from which to try recovery of body weight. Hauger (1921), for instance, has experimented with a quantity obtained from the combined volume of right humerus, radius, ulna, femur, tibia, and fibula. Using Australians, Europeans, and a scattering of others, he has obtained results that parallel what he might have gotten from body weights. However, this sort of thing has the vice of its own virtue. As a technical matter, it demands that the collector of fresh material clean all of these bones. Often that is highly impracticable.

Accordingly, we have dealt only with brain and body weights. Some data do exist. They may be far from uniform in quality or quantity, but one thing is certain: when they are put through a process, they do tell a story.

B. THE RANGE OF THE MATERIALS

1. Perspective

They are fetal and postnatal. I have tabulated raw data, straight through from the earliest obtainable up to, and (sometimes, but not always) including, adults. It will develop that, in each case-subject, the entire stretch available for a fetal period could be used for a single formula.

However, in the postnatal life, only those after some period of "infancy" could be used safely in a single formula. Accordingly, I have formulated growth curves for a life-stretch that is "postinfantile," and which terminates at adulthood. Between birth and this last period, all animals trace transitional growth curves which demand special treatment. Where to begin and where to break off is no certain matter. (For safety's sake, I have formulated the human "postinfantile" stretch from data not earlier than the age of 4. For the macaque, I have included the stretch from the beginning of permanent-

tooth eruption to its completion, but not beyond. This difference of standard has no perceptible effect upon the formulations of the curves.)

The tables, then, will include some data which were not used in the formulae. Nevertheless, usually they have been plotted in the figures The reason will be obvious on each occasion.

2. Index of Data

(1) ***Homo***	(a)	*Fetal:* 1. Composite series from Cruikshank and Miller (1924), Giese, and Rüdinger (both *fide* Cruikshank and Miller).
		2. The averages of Michaelis (1906).
	(b)	*Postnatal:* 1 Composite of Michaelis (1906) and Wendt (1909)
		2. Collaterally, tables 8 and 27 of E Boyd (1942).
(2) ***Pan***		*Postnatal:* Data from various sources assembled by Anthony (1928), including his own.
(3) ***Simia***		*Postnatal:* Composite of Anthony (1928) and Hrdlicka (1925).
(4) ***Rhesus***		*Postnatal:* Zuckerman and Fischer (1927).
(5) ***Semnopithecus***	(a)	*Fetal: S. maurus*—Hulshoff-Pol, *fide* Anthony (1928).
	(b)	*Postnatal: S. obscurus*—Keith (1895).
(6) ***Felis***	(a)	*Fetal:* H. B. Latimer, unpublished.
	(b)	*Postnatal:* Ziehen (1901); also, Anthony (1928).
(7) ***Mus***		*Postnatal:* Donaldson and Hatai (1911).
(8) ***Ovis***	(a)	*Fetal:* Ziehen (1906).
	(b)	*Postnatal:* Crile and Quiring (1940).
(9) ***Bos***	(a)	*Fetal:* Ziehen (1906).
	(b)	*Postnatal:* Crile and Quiring (1940).
(10) ***Sus***		*Fetal:* Ziehen (1906); also, L. G. Lowry, *fide* Anthony (1928).
(11) ***Equus***		*Postnatal:* Crile and Quiring (1940).

3. Index of Tables

Brain and Body weights of:

TABLE

4. Notes on the Data

TABLE 1. *Homo.* Fetal: Michaelis' fetal averages probably are too low in body weight for normal fetuses. Thus, at 9 months, he gives 1994 gm. A European infant weighs ½ again as much and even more at birth. The composite of three authors for the other human formulation seemed justified from a plotting.

Postnatal: Michaelis' and Wendt's data overlap excellently on plotting. It is uncertain how early one may initiate the series for a straight-line logarithmic formula. To be safe, nothing under the age of 4 years was used. This probably erred on the side of scrupulosity.

TABLE 3. *Orang.* Only male specimens from Anthony were used for formulae, because there is only one immature female, and adult sex dichotomy of size is very marked. Hrdlicka's are all mature specimens. They were not used, because their weight would have vitiated the calculation.

TABLE 4. *Semnopithecus.* The two species represented here are not far apart in their adult proportions. At any rate, one may doubt

that their growth behaviors differ too much to furnish a notion of growth in this genus.

Not tabled: *Rhesus*. Only data from beginning to recent completion of permanent tooth eruption were used for calculations

TABLE 5. *Felis*. Anthony (1928) was checked against Ziehen (1901) The postinfantile straight line begins fairly with the 2-month age, regardless of the matter of dentition. (Correlation between onset of permanent tooth eruption and the straightening-out of the line of growth we are studying, is a problem for future investigators.)

Not tabled: *Mus*. Donaldson (1924) *computes* brain and body weights of Norways for every millimeter of body length, treating the sexes separately. The formulae he uses are Hatai's (1909):

$$Br = 0.569 \log (Bw - 8.7) + 0.554, \text{ when } 10 < Bw < 325$$
$$Br = 1.56 \log Bw - 0.87, \text{ when } 5 < Bw < 10$$

Together,

$$Br = 0.825 \log (Bw - 4) + 0.233,$$

sexes combined; then a correction up and down is made for the sexes These formulae are very completely empirical. They are hardly adapted for our purposes, however successful they may be for predicting brain weight from body weight.

I therefore reverted to Donaldson and Hatai (1911), for the form in which they presented the data upon which their formulae are based. That form is one of grouped data. I have disregarded the albinos, because I wished to avoid the complications from domestication wherever possible (for the artiodactyls, the data had to be from domesticated forms). For the straight-line formulae, only the data between body weights 15 and 485 gm., inclusive, were used.

TABLE 6. *Ovis*. Only the last three entries were used. They are considerably older than the first two, which, respectively, are fetuses and 1-month-olds.

TABLE 7. *Bos*. In the postnatals, only entries 3 through 7 were used. The first 2 entries are too young to fit the straight-line post-infantile formula, and entry 8 is the sole female among the immatures. Hence, the entries used represent a male growth curve only.

TABLE 8. *Sus*. The entries represent means of lustra. The period of prenatal growth so recorded is exceptionally long.

TABLE 9. *Equus*. On these principles, only the starred entries were used for formulation.

C. RELIABILITY OF THE FORMULAE

TABLES 11 and 12.

If even a small series happens to form a good straight line, it favors reliability. If the points of postinfantile growth are spaced over a considerable distance, they favor an accurate formula more than if they be bunched near one end of the putative line. A large population is generally more reliable than a small one, but I know of no satisfactory formula for calculating the P.E. of a line.

In the matter of sex difference, the human values of TABLE 11 are very close together; enough to raise the possibility that the difference is not significant. A collateral experiment on figures of E. Boyd (1942) (see below, under D1, *Man*) reverses the position of the sexes, in point of relative size of the constants. However, the brain-body ratios are never alike in the two sexes; this is known empirically. So, while they are known to be different in the sexes at various ages, we cannot be sure, from our formulae, which grows faster in the brain/body ratio, the male or the female.

The anthropoid ape data are meager, yet, together, they agree closely (TABLE 11).

Sex discrepancy in *Rhesus* (TABLE 11) probably exists, but the *size* of the discrepancy is at least striking.

For *Semnopithecus*, the data on the immature are few, the line is short, the few adults probably deflect the line Perhaps a value of .14+ for the *b*-constant would be more nearly truthful On the other hand, the sexes in *Rhesus* deviate considerably, and cephalization in the Cercopithecidae varies strongly. So, perhaps, the difference in growth-ratios also, between *Rhesus* and *Semnopithecus*, is real enough.

It would take but very few more fetuses, properly placed, to swing the fetal slope in alignment with the human (FIGURE 4).

The alignment of the postnatal cat data is fairly good. The prenatal data are peerless.

The rat data are excellent, so that the sex difference is probably genuine.

The materials for the postnatal Bovidae and the horse are not abundant, but the agreement among them (TABLE 11) is excellent. It points to *sub*equality, which is what one might look for. The prenatal pigs cover an exceptionally long period of growth. The entries represent averages of groups, and the number of samples is very unusual. But the data do not line up very well, although a straight-line general trend is acceptable. The point will be brought up again later.

D. BEHAVIOR OF THE GROWTH CURVES

1. Man

FIGURES 1–4, 6. TABLES 1, 11, 12.

The mass tendency of growth of brain and body weight in man and woman, from the earliest fetal stage available, describes a steep straight line when plotted logarithmically. At some indeterminate period, the line very gradually deviates from the straight. Birth occurs while the deviation is still too slight to be ordinarily perceptible. Thereafter, during infancy, the curvature becomes marked. More abrupt than the fetal deviation is its convergence upon, and merging into, another straight line, which characterizes growth, thenceforward. Just when this new straight line becomes valid, is also indeterminate; but, for practical purposes, it has safely supravened by age 4 years

Since our fetal data have not warranted separating the sexes, the line we have drawn is a sex compromise. At best, it could hardly be expected to join perfectly the postnatal constellations of the sexes.

The rest of the procedure is graphed in FIGURE 3 (see TABLES 11 and 12) which gives velocities of growth for each sex, assuming the fetal velocity as about 1.0098.

Mathematically,

$$\frac{dY}{dX} = b,$$

whenever $Y = A + bX$. Let b' be fetal velocity, b'' postnatal velocity. Then:

$$♂ : b' - b'' = 1.0098 - .06326 = .96874$$
$$♀ : b' - b'' = 1.0098 - .07010 = .89399.$$

In FIGURE 3, the values of b' and b'' are asymptotes. The S-curve is a skew logistic, and its integral is the growth behavior transitional between fetal and the established postnatal As far as I know, this integral is unsolved. However, fair values of Y' can be obtained from:

$$♂ : Y = \frac{.96874}{1 + \text{antilog}_{10}\,(1.6922X^3 - 18.3695X^2 + 67.6053X - 84.5844} + .06326$$

$$♀ : Y = \frac{.89399}{1 + \text{antilog}_{10}\,(-1.850X^3 + 17.980X^2 - 56.934X + 58.151} + .0701.$$

The essentials of our description are simply these, and they could have been immediately read out of the curves of FIGURES 1 and 2:

(1) The transitional period is marked by a velocity which is asymmetrically S-shaped;

(2) The adjustment to postinfantile growth rate, therefore, is much more abrupt than the deviation away from fetal behavior;

(3) For the data given, it differs quantitatively in the sexes.

But what lies back of this manifestation?

Donaldson (1917, p. 136) says:

"At birth cell division is still in progress, and this is true for both man and the rat; moreover, for some time after birth cell division continues—especially in the cerebellum. Allen (1912) finds in the *cerebrum* of the rat some cell division up to the twenty-fifth day of postnatal life, after which time it diminishes rapidly and soon becomes insignificant. In the cerebellum, however, postnatal cell division is more abundant than in the cerebrum and is responsible for considerable change. The enumerations are given in TABLE 3.

"TABLE 3

BRAIN OF ALBINO RAT MITOSES IN ONE CUBIC MILLIMETER OF NERVE TISSUE (ALLEN, 1912)

Age days	No of mitoses—Brain	
	Cerebellum	Cerebrum
1	1597	430
4	2111	447
6		193
7	4848	
12	839	37
20	127	23
25	00	27"

Whether coincidentally, or whether in terms of cause and effect, the cessation of neural mitoses, apparently variable locally, occurs during the transitional period. The steep fetal velocity covers a period of growth in weight, plus number of cells in the brain; the gentle postinfantile velocity covers a period of growth in size of cells, without increase in number. However, much of the rest of the body continues, in the latter period, to increase in terms of both cell multiplication and cell size.

A glance at FIGURES 5–9 will indicate that man is not peculiar in these matters.

2. Man and Other Primates

FIGURE 4. TABLES 1–4, 11, 12.

At face value, the human fetal slope is the steepest; the postinfantile, one of the gentlest. In the matter of the fetus, the monkey data are too few for a positive conclusion. (See also below, under "Cats.")

On the postinfantile side, the story is clear: Monkey brains grow at the *steeper* rate; man and anthropoid stand off together with *lower* rates. In fact, the anthropoids have a lower rate than man. Consequently, the rate does not correlate perfectly with relative brain weight.

Moreover, the fetal rate (*i.e.*, over the distance for which we have data) does not account for the great preponderance of the human brain weight and the "second-prize" weight of the anthropoid. For the human and *Semnopithecus*, rates are *subequal*, anyway, and the two chimpanzee entries are consistent with the scheme. Thus, an average between the human formula from Michaelis and that from the composite series (TABLE 12) is $Y = -.7345 + .9619X$ If this, in turn, be averaged in with *Semnopithecus*, we have a rough general trend of $Y = -.64575 + .8847X$. The two chimpanzee entries are joined by a line which happens to parallel it exactly. A slope of something in the neighborhood of .9 should be a fair center of approximation for Primates. It is very close, we see, to 1.0; —tan 45°, where increase in brain logarithm would exactly pace increase in body logarithm.

While the rate may be one factor, then, if human (and chimpanzee) rates are at all higher than monkey, it must be less important than some other. Notice that, at equal body weights the human brain is heavier than the simian; a larger percentage of the human body is brain when total body weights of the fetus are equal This ascendancy had to be achieved at some age earlier than those registered on the chart. Further than this, at a given body weight equal in man and monkey, the monkey is the closer to birth. When the monkey is being born, thereafter to enter upon the transitional deviation that leads to the side-line position of the adult, the human, still underdeveloped by comparison, is continuing to travel up the fetal curve. The achievement of the human *proportion, as well as size,* of brain *demands an enlarged body*. It looks as though the human, to achieve his high cephalization, shifted the mutual relationship between two blocks: the energy of growth which determines size, and the urge toward definition which determines morphological maturation.

In Section IV, we shall see that the requisite of an enlarged body, as a *sine qua non* of a proportionately larger brain, is borne out by phylogenetic considerations.

The postinfantile ape behavior will no doubt raise anew the question of whether the difference between the stated human growth rate and the anthropoid is genuine or spurious. It says that, although the monkey brain, after infancy, grows at a *higher* rate per body than the human, the human grows at a higher rate than anthropoid.

It is true that the human data all represent natural deaths, but so, for that matter, must at least a goodly share of the anthropoid. Presumably, the bodies were more or less underweight. Then the body weights would be too low per brain weights. Would the underweight condition be so distributed along the entire line as to alter, or not to alter, its slope? I do not know. For the sake of discussion, let us assume, first, that the human line should more nearly approximate the anthropoid, or even be lower. As an experiment, then, I plotted the brain weights in E. Boyd's (1942) Table 27 against her body weights in Table 8 (Minneapolis and Chicago children). Of course, the brains did not belong to these children. At least, this should correct for body weight deficiency, though not for brain weight deficiency (which probably is less discrepant), and we might look for a lower slope, after the doubt just raised concerning the condition in man. Here are the results:

$$♂ : Y = 2.8215 + .06859X \qquad y = 663.0x^{06859}$$
$$♀ : Y = 2.8701 + 05002X \qquad y = 741.4x^{05002}.$$

This reverses the sex behavior of the slopes, as compared with the formulae of TABLE 11; still, the male formulae are practically identical. At any rate, the new slopes differ remarkably little from the old. The behavior, therefore, seems genuine. Then let us reverse the assumption. Should not the ape slope be much steeper—intermediate between man and monkey? In answer, chimpanzee and orang behave almost identically. Thus, their data become reliable, in spite of their paucity.

Nevertheless, it is obvious from the chart that the several brain-body positions of man, ape, and monkey can be achieved without demanding the straitjacket of an *a priori* consistency. What we are witnessing, after mitosis has ceased, is increase in brain weight divorced from the additional factor of cell multiplication. Not being a neurologist, I prefer to waive further comments on this phenomenon. It is enough that we have seen that brain preponderance is a function of the fetus. However, it must be that the possible factor of steeper increase in brain with respect to body is practically over with by the time our measurable period begins; for thereafter, another factor is manifest: brain preponderance is in terms of a steeper increase in *brain-plus-body-weight per level of morphological maturation.*

3. Man, Other Primates, and Some Ungulates

FIGURES 5, 6. TABLES 6–9, 11, 12.

a. As compared with the primates, the cattle show:

(1) A gentler fetal slope (average for the three together: .7801, as compared with the primate, .8848).

(2) Postinfantile slopes on a par with those of the monkeys, but steeper than those of man and the apes.

(3) About the same range of fetal slopes as that from monkey to man. If, in TABLE 12, we subtract one b-value from another, we obtain:

Cattle:	
Sus-Ovis	.1372
Sus-Bos	.2562
Ovis-Bos	.119
Homo-Semnopithecus:	
Using the composite human formula:	.2015
Using Michaelis' data:	.1065
Using the mean of these two:	.1540

But this may signify very little.

(4) A marked equality or subequality in the magnitude of the postinfantile slope; whereas the Primates vary, from monkey to anthropoid. This should be noted in conjunction with the taxonomic ranges represented, which are, respectively, Hominidae, Anthropoidea, Cercopithecidae, and Suidae, Bovidae among artiodactyls, and, in addition, two of the Equidae.

(5) What hardly appears from the chart, yet is necessarily of a piece with the rest of the traits, is that the range of *cephalization* among these ungulates is very narrow; it is very wide among the primates.

Cephalization is a term to be exploited in the section on phylogeny. Just now, it is convenient to define it as:

$$\text{Cephalization} = k = \frac{(\text{Brain weight})}{(\text{Body weight})^{56}}$$

The values for our ungulates and primates are:

Equus	.43		
Bos	.30		
Ovis	.32		
Sus	.17		
Range:			.26
Cercopithecidae (without baboons)	1.11		
Simia	1.21—	.75	
Homo	2.75		
Range:		1.64—	2.00

The situation reduces to this:

Primates. Man and monkey have subequal fetal slopes. Man climbs that slope (using the earliest periods registered on the chart, and disregarding the unknown situation anterior to them) along a slightly higher elevation (per body weight). More notably, he climbs it much farther before turning aside to the right. After he turns, he travels but a short distance, and on a much more gradual rise than the monkey. As a result, he is bigger-bodied, bigger-brained, more highly cephalized.

Artiodactyls. Cattle have subequal fetal slopes; these slopes are slightly more gradual than the Primate; no one of them climbs along a markedly higher level (per same body weight) than another. *Bos* climbs this slope farther than *Ovis* or *Sus,* before turning aside to the right. After turning, all three travel on a rise of the same general magnitude as that of *Rhesus* or, for that matter, of *Felis.* (See TABLE 11.) Result: *Bos* is bigger-bodied, bigger-brained, but no more highly cephalized than *Ovis; Ovis* is smaller bodied, slightly bigger-brained, and, along with *Bos,* a little more cephalized than *Sus.*

That it takes a number of metric factors to bring on the end-product called an adult, should be clear; also, that these factors can vary in a number of combinations; but further, that the variety is not random, for the combination has taxonomic value. Shortly, we shall be exploiting those devices of Dubois, the relational exponent and the cephalization coefficient. They have been useful for several decades, and the criticisms of their shortcomings (which are real) have not destroyed them. But if we expect to use them, it is very advisable to go behind the scenes before these actors take the stage.

b. What then, is the summary of factors by which the mean adult reaches his final position on the chart?

(1) The initial position of his fetal slope;
(2) The degree of that slope;
(3) The level at which he leaves that slope and turns to the right;
(4) The length of the transitional curve;
(5) The curvature of that curve;
(6) The degree of the postinfantile rectilinear slope;
(7) The distance traveled along that slope.

Factors 2 and 6, taken together, form an angle which is smaller, the greater the stepdown in velocity from the fetal level to the postinfantile. The anthropoids and man have the smallest angle; in fact, the anthropoid drop in velocity is even greater than in man. Yet, the

anthropoid is only a "second-prize" brain. So this compound factor must be taken in conjunction with another: the level at which the postnatal line "peels off" to turn to the right. As a practical measure, it is best indicated by taking the point at which the projected fetal straight line would intersect the projected postinfantile. This may be achieved by solving the two formulae as a pair of simultaneous equations. In short, the location of any total one of the developmental lines may be viewed as essentially a pair of straight lines forming an angle at some point in a Cartesian field and in terms of three quantities: the velocity or slope of *either one* of the two straight lines; the point X, Y where the two intersect; and the difference between the fetal and the postinfantile velocities This leaves the beginning of the fetal line indeterminate, but, for present purposes, inessential. The mean point of adulthood still remains to be determined, and this is essential.

4. Cats and the Law of Allometry

We broaden the comparison by considering FIGURE 8. No fetal formula appears in TABLE 12, simply because Latimer's data really are *too* good. Now we can see definitely a case of something suspected hitherto: the fetal line is not really rectilinear at all (at least it is obvious in the cat), but considerably more complex. The total cat path, from its earliest record to a projected merging with the postinfantile period (and perhaps even that should be included), may be a parabola of no less than third degree. It gets steeper, as it approaches birth, and a point of inflection in the line is suggested at that period. There may, also, be one near the lower or earliest end of the constellation.

If the other two long fetal lines that we have (Michaelis' human averages, and the pig) be scrutinized closely, a similar possibility will appear, though it is less pronounced. Being less pronounced, it takes a rectilinear fit fairly well. So, as we have been comparing slopes of Bovidae and Primates, we have overlooked the possibility that we should do so only with those portions of the line that are strictly comparable. But what is the unit by which they could be compared?

It should be that of logarithmic cycles. Thus, data on the brain of the fetal pig are given for a little over three logarithm cycles; the cat's covers about 2½ cycles; Michaelis' human averages, a little less than 2 cycles. Viewing these lines as curves having at least one arc, the deepest place in the arc occurs at about the onset of the last prenatal logarithmic cycle. That region is very clear in the cat; it is fairly so in Michaelis' humans; the pig is less certain. If the lines are really curvilinear, it is readily understandable not only why, within an order,

the pieces of slope have varied as much as they have, but also why the values of the A-constant differ so widely.

Since, today, there is a widespread tendency to apply the "law of allometry" ($y = ax^b$) to all sorts of growth behavior, it is timely to caution that the law often may be but a first approximation, pending the more accurate application of complex exponents in place of the b in $y = ax^b$. If the "law of allometry" be adopted too wholeheartedly, it can turn obscurantist. For, if undulations actually be present in a line of growth through which we have run a rectilinear shortcut (and "shortcut" is exactly what it would be, if the variable y keeps changing its rate of increase, with respect to that of x, by first accelerating, then decelerating), they are real phenomena which that formula screens from further investigation. If our fetal growths have been really undulatory, all the time, then we have been comparing pieces of arc without first standardizing their lengths. This does not mean we have been wasting effort: an undulatory set of growth curves would still have to be analyzed along the lines laid down here.

These reservations could be easily illustrated, if necessary, by way of the artiodactyls, but lack of space prevents this. Considered this way, the apparent failure of the fetal and postnatal lines of human ontogeny to join perfectly (FIGURES 1 and 2) may not be due entirely to individualisms of technique in the collectors (although some of both the fetal and postnatal data are Michaelis', which is one reason for using his material), but may reflect, also, a genuine condition, one resembling the feline.

SECTION III. COMPARATIVE ANATOMY

PART ONE

RANGE OF MATERIALS: ADULTS

1. Index of Data

1. *Ungulates.* Brummelkamp's (1939c) Table I. See TABLE 13.
2. *Rodents.* Brummelkamp's (1939b) Table I. See TABLE 14.
3. *Carnivores.* Crile and Quiring (1940); Hrdlicka (1905).
4. *Primates.* Spitzka (1903), averages for the species mine; Hrdlicka (1925), his averages corrected. *Homo:* Crile and Quiring (1940). (My average from all adults.)
5. *Other Orders.* Crile and Quiring (1940); Hrdlicka (1905).
6. *Palaeanthropoi.* Hrdlicka (1930), his own data and others'.
7. *Reptiles.* A: Crile and Quiring (1940); B: Dubois (1913).
8. *Amphibia.* Brummelkamp's (1939d) Table I, from Dubois (1913) and Donaldson (1910).
9. *Weights of Human Brain by Parts.* Marshall (1892).

2. Index of Tables

TABLE

PART TWO

PRELIMINARY

a. Hitherto, as described in the Introduction, we have presupposed that a line of mean tendency in growth was a sort of summary description of the path actually traveled by the average, or standardized, individual as he grew. That it is an assumption, should be emphasized now, for the last time. As a practical matter, the measurements on fetuses have been taken at one moment of growth in each individual. This cannot be simply assumed to be the same thing as measuring the same individuals over a successive period of time.

We propose now to compare the adults of extant genera, and to do so by orders. We ask, Do the members of an order group themselves in such a way as to give a definite and consistent scheme?

This is an appeal to comparative anatomy. It would be very desir-

able to be in a position to pass from these modern animals to their phylogenetic antecedents. Some day, this will be possible; at this juncture, it is not. All that we can say is this: If it should prove that modern members of a common order plot in a regular fashion that is even mathematically formulable, we can be morally confident that the order itself has differentiated in no haphazard way.

Technically, then, we shall seek to formulate curves, as before; and, since such curves can represent only connections between extant contemporaries, let us call them, for want of something better, "*comparative anatomy*" curves. The true phylogenetic curve would connect a series of direct evolutionary ancestors. It would lead up through a succession of fossils, located in the time of previous geological epochs, to present times and extant forms If, for a moment, we glance at almost any of the graphs that follow, we may picture the surface of our sheet as located in the time of today, lying on an indefinite pile of such sheets, each charting the condition of the animals extant in some previous epoch. The true phylogenetic line should wander *up* through these sheets, piercing each in turn, until it emerges on the surface now exposed to us.

Immediately, it will be seen that time is the third dimension, its axis being at a right angle with the plane of the paper we are looking at. The true phylogenetic curve, in other words, is one of three coordinates, x, y, z, in which x and y remain the measures, respectively, of body and of brain weight (taken as logarithms X and Y), while z is a time-scale.

Suppose, next, we imagine all our sheets to be transparent. Then (confining our example to the primates), on the top sheet we may perhaps draw a curve connecting the modern primates with each other: the comparative anatomy curve. Gazing straight down through our pile, now, we see true phylogenetic curves coming up at us. How closely, however, does our pseudo-phylogenetic curve appear to cover, or project its shadow upon, the true phylogenetic line? More technically, how closely does our comparative anatomy line coincide with the projection of the true line onto the x, y plane?

This is what we cannot answer, as yet, for want of sufficient fossils. This is why we must content ourselves with plotting the curve that lies on our top sheet alone. This means tracing curves on the basis of com parative anatomy. At least the principles to be laid down anticipatorily should then need merely to be expanded into the geometry of three dimensions, whenever the fossil record becomes adequate.

b. The comparative-anatomic path actually has been treated by some authors as though it were the true phylogenetic path, particularly

since Dubois introduced the view that the brain weight of any mammal (eventually, of any vertebrate) is related to its body weight by a formula $y = ax^b$ (1897). (Dubois used a different set of symbols. See below.) He has been followed and supported by Lapicque, Anthony, Ariens-Kappers, Brummelkamp, and others (see Bibliography).

The mathematical law, $y = ax^b$, has found widespread use in other biological investigations, under the name of "law of allometry," especially since the impetus given it by J. Huxley (*Problems of Relative Growth*, 1932; see the bibliography in Hersh, 1941). In the literature on brain/body-weight, it has been variously written $E = kP^r$, $E = cP^r$, $E = kS^r$, etc., so that, in the following development, any or every one of these forms will have to be used, depending upon which author we are discussing.

A. THE "PHYLOGENY" OF BRAIN/BODY WEIGHT EXPRESSED AS $y = ax^b$

1. The Relational ("Interspecific" or "Phylogenetic") Exponent and the "Cephalization Coefficient"

In 1897, Dubois, assuming that certain animals, closely related, but differing widely in body size, must have equal cerebral organization, propounded that their brain weights would be to each other as the rth power of their body weights:

$$E : e = P^r : p^r \qquad (1)$$

where E and e are weight of "*encephalon*," P and p are body weight, the capitals represent the bigger animal of a comparable pair, and the minuscules the smaller animal. Dubois initially compared the following animals:

Simia satyrus — *Hylobates syndactylus*
Simia satyrus — *Hylobates leuciscus*
Oryx beisa — *Cephalophus maxwelli*
Felis concolor — *Felis domestica*
Felis leo — *Felis domestica*
Mus decumanus — *Mus musculus*
Sciurus bicolor — *Sciurus vulgaris*

E, e and P, p being measured and therefore known, r can be found:

$$r = \frac{\log E - \log e}{\log P - \log p} \qquad (2)$$

In every case tested, the values of r *approximated* .56. Even with many further pairings, including all or most classes of vertebrates, the quantity generally varied within what seemed to the investigators a

reasonably narrow range. Hence, r, which, when we write $y = ax^b$, is b, has been termed the "*relational exponent*." (See Dubois, in Bibliography.)

Once $r = .56$ is taken as universally valid, the brain weight, y, of any animal is related to its body weight, x, as:

$$E = kP^r \text{ or}$$
$$y = ax^b,$$

where $r = b = .56$. In this formula, if y and x have been measured, a can be calculated:

$$a = \frac{y}{x^b},$$

and a (or k, above) is the "coefficient of cephalization." If, for instance, we have two animals of practically identical body size (say gorilla and lion), but obviously of discrepant brain size, then the values of a will differ, because their y differs; and the size of a favors the gorilla, who is more "cephalized." Conversely, then, if a has been ascertained for various kinds of animals, and the weight x of an animal be known, $y = ax^{.56}$ gives approximately the brain weight. Where two animals are equally cephalized, but differ in brain and body weights (*e.g.*, lion and cat), they will actually be related (verting now to the Dubois symbols) as:

$$E : e = k\,P^{.56} : k\,p^{.56} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (3)$$

but k being identical for both, the formula reduces to EQUATION 1.

At the present date, those individuals who are satisfied that intelligence is not a correlate of brain mass, will discredit the coefficient completely. On the other hand, those who are not ready to go so far and who believe that mass is still one correlate, even if only an imperfect one, of cerebral capacity (whatever be the definition of that term), will insist that the relative size of brain in gorilla and lion is not purely coincidental or insignificant. At least, the presuppositions underlying the work of Dubois and his followers, at the close of the last century, are no longer tenable. At that time, it was hoped that an anatomical separation might be effected between the "psychic" brain and the "somatic"; the former controlling the "higher" mental functions, the latter, the physiological ones. Von Bonin, in fact, apparently rejects the Dubois scheme, primarily because of this naïve conception (1937). Criticisms, however, must lie in abeyance until after we have examined the scheme more fully.

The formula $y = ax^b$ is grasped better if it be pictured. On a log-log scale, it turns into $Y = A + bX$, where $Y = \log_{10}y$, $A = \log_{10}a$, $X = \log_{10}x$. Thereby, b becomes the slope of a straight line, and A is the

value of Y, when $X = 0$. In such a disposition, the pairs of animals of Dubois should be united *intra se* by sloping, parallel bars which, when projected to the Y axis, should mark values of A (FIGURES 10, 11, 13, 14).

2. The Ontogenetic Exponent

One might be led to expect that, if cats and tigers are related as

$$E : e = P^{.56} : p^{.56},$$

then different-sized domestic dogs would most certainly be so related. Certainly they are in the main equally "cephalized." But no; they are related as

$$E : e = P^{.25} : p^{.25},$$

.25 being an average figure obtained from numerous pairings as before This finding is Lapicque's (1898, 1907, 1908, 1912). The constant $r = .25$ has been termed the "ontogenetic" or "intraspecific" exponent.

Why animals within a species should so behave, has led to speculations (*e.g.*, Dubois 1923). Whatever the explanation, the Dubois school has built further upon it. One may see its use by Anthony (1928), who extensively compares animal brain weights; or Harris' experimental use of it (1929); also Ariens-Kappers (1929, 1936). Anthony and Coupin (1925, 1925–26, 1929) have erected out of it an "Index of Cerebral Value," which shall be discussed a little further on.

3. Discussion of the Dubois System

The rationale of the exponent or slope .56, to my knowledge, has never been explained. That all the vertebrates compared by Dubois and his followers give values that seem to them satisfactorily close in agreement, has naturally led them to conclude that there is something fundamental in, and common to, all the vertebrates, which is masked by this number.

However, the diverse values of the cephalization coefficients have led to other speculations.

At the outset, Dubois' comparisons of animals having different values of y, x, and a are not strictly and explicitly phylogenetic. A cat and a lion are two modern Felidae. Their mathematical relationship is one existing today. The one animal is not derived from the other. They have had a common ancestor which, whatever else may have been true of it, could not have equaled both of them in body and brain size. The fact that both have the same cephalization coefficient may be explained as one will. Either it was the possession of their common ancestor, or

else both lion and cat have arrived at the same value by a parallelism. If the former be the case, then the bodily differentiation in point of size has proceeded without any change in cephalization, and we have in the cephalization a now static situation. If the other be the case, then cephalization has been an identical and concurrent achievement, even while brain and body size were differentiating. This would lead to the conclusion that there has been an identical potential towards cephalization which inhered in the common ancestor.

These are the speculations which must suggest themselves to the student who is so evolutionally-minded as to insist upon raising to the surface the implications of the Dubois system. The Dubois school has recognized that there must be implications, although it has not expressed them in such forms as above. Instead, Dubois and his followers have sought to explain the differences in the cephalization coefficients in terms of quantitative cytoarchitectonics. The most recent studies are those of Brummelkamp (1939–40; see also Dubois, 1913) He devises a scale for the cephalization coefficient (which he writes as c), such that:

$$c = \sqrt{2}^{a} \text{ or } \log c = a \log \sqrt{2}, \qquad \text{where } a = 0, 1, 2, 3 \ . \ . \ .$$

(See our FIGURES 13, 14.) Basing his theory upon Dubois (1913 *et seq.*), he rationalizes this in terms of cytoarchitectonics However, into this aspect of the question we shall not delve The points that are relevant here are:

(1) Brummelkamp makes evolutionary implications practically inevitable, by concluding that, within any order of animals (*e.g.*, rodents or ungulates), cephalization increases saltatorily in units of $\sqrt{2}^{a}$, as one passes from the less cephalized of the order to the more cephalized

(2) Brummelkamp's own graphs indicate that greater cephalization accompanies increased body size An aspect of the phenomenon that is crucial to the development of the system to be elaborated in the present essay, but which is *not* explicit and unalienable in the Brummelkamp system, is the possibility that the different stages of body size, plus their relative amounts of brain, as illustrated in FIGURES 13 and 14, correspond, in general, to the path actually taken by the order in its quantitative cerebral development. This, then, is the issue: Can animals within a given order, having small bodies and a "lower cephalization," be considered at all as "contemporary ancestors"? To this we must return later on. Meanwhile, now that the character of the system of the Dubois school is before us, we may turn to some critical analysis.

4. Criticism of the Dubois System, and of Some Derivatives from It

a. The fact that the scheme is built on a naïve philosophy of the brain has already been mentioned. This, none the less, would not overthrow a set of empirical findings. If the Dubois and Lapicque equations really hold, then they are empirically valid, however they may be explained. But do they hold?

b. The method of ascertaining values of r or of k was one of simple algebra, and one of pairing together animals assumed to have equal "psychic" endowments. This would not be considered good statistical technique today. Then, can one reach the same results by using modern methods? As will develop, one can do so only within certain limits which exclude many contradicting cases, notoriously among the Primates. In fact, it will transpire that Dubois' obtaining $r = .56$ out of comparisons between orangs and gibbons was only a fortunate accident, and one, furthermore, that is its own refutation. No cerebral anatomist would consider the brains of gibbon and orang at all comparable in the sense of *Felis leo* and *Felis domestica* or of the two *Mures* or *Sciuri*.

c. This weakness is critical, not only in the matter of $r = .56$, but in that of $r = .25$. A closer examination of Lapicque's use of his dog material throws an additional doubt upon the scheme.

Lapicque grouped his animals into ten body sizes and weighed brains and bodies. Then, by comparing group I with group II, II with III, etc., he obtained a mean r of approximately .25.

The question now arises: What about man? Should not the relation between male and female run according to $r = .25$, assuming that they are equally intelligent, members of the same species, and different in body size?

Lapicque was writing in the early part of the twentieth century (1907) when the "emancipation of women" demanded from the highest court of appeal (that of the scientist) a clear proof that women were men's "intellectual equals." Lapicque evidently believed he had supplied it. However, we do not mean to imply that he produced his proof on a direct and explicit demand. Here are Lapicque's operations:

$$\begin{aligned}
♂ : P^r &= 60\,000^{.56} \text{ gm.} = && 498 \\
E &= 1\,360 \text{ gm.} \\
k &= 1\,360 \div 498 = && \mathit{2.73} \\
♀ : P^r &= 54\,000^{.56} \text{ gm.} = && 448 \\
E &= 1\,220 \text{ gm.} \\
k &= 1\,220 \div 448 = && \mathit{2.72}
\end{aligned}$$

From the values of k, he concludes, "*Il y a égalité.*"

Probably the most striking trait of this arithmetic is the use of $r = .56$, instead of $r = .25$. Lapicque insisted that men and women were so unlike, somatotypically, as to make the difference one of generic rank. Whether this difference is more marked than those between a mastiff, a greyhound, a dachshund, and a pekinese, he does not say. Anyway, were we to use Lapicque's own values of P and E, and the exponent $r = .25$, we should obtain:

$$♂\ k = 8.48$$
$$♀\ k = 8.02$$

Voilà qu'il n'y a pas d' égalité

We confess a surmise that it was the failure to obtain equality here that led to Lapicque's use of $r = .56$ But if we fall in step with him and use the data from Welcker's (1903) 4 executed males and 3 females, we obtain:

$$♂\ k = \frac{1461}{59270^{.56}} = 3.18$$
$$♀\ k = \frac{1249}{50800^{.56}} = 2.89$$

From Crile and Quiring's (1940) 32 males and 9 females, of assorted races, we obtain:

$$♂\ k = \frac{1350}{65544^{.56}} = 2.70$$
$$♀\ k = \frac{1214}{49755^{.56}} = 2.86$$

It seems, therefore, that Lapicque struck his "*égalité*" by accident. If, on the other hand, we stop assuming $r = .56$ and by a modern method seek the empirical equation for the human situation from the raw data of both sexes, we obtain,

From Lapicque's figures: $Y = .5255 + .541X$
$y = 3.35x^{.541}$

From Welcker's: $Y = -2.4053 + 1.167X$
$y = .03933x^{1.167}$

From Crile and Quiring's: $Y = 1.2773 + .385X$
$y = 18.94x^{.385}$

The discrepancies speak for themselves.

Returning to Lapicque's dogs and his intraspecific exponent, FIGURE 12 should illustrate the reason why his r is so much less than .56. The adult dogs, each presupposing an ontogenetic line such as that already developed in Section III, align themselves definitively so that the axis of their adult scatter inclines at a slope of .23 (using Lapicque's own data, and solving by the method of weighted-averages). Morpholog-

ically interpreted, small domesticated dogs have unduly large brains, and *vice versa,* when a slope of .56 is used as a standard of reference. If we had the individual measurements, we could find a correlation coefficient by modern statistical methods, and we might find that man's selective breeding which has sorted dogs into sizes and strains, has "infantilized" or dwarfed some and has "feralized" others, or increased their body size without correspondingly pushing up their brain size. This would be a case of "stretching" the species along the X-axis of a graph, without distorting in equal proportion the Y-axis, a phenomenon which is handled in analytical geometry under the term "transformation of axes." (This concept will prove very useful later on.) Unfortunately, we do not have the requisite data on wild Canidae to verify this by way of a proper, and a very interesting, comparison.

And if man were to transfer the breeding and domestication to cats and tigers, he might well find "ontogenetic" slopes in each. Yet, the *dead center* of the cat population might still connect with that of the tigers by way of a slope of .56.

d. The next criticism concerns the "Index of Cerebral Value" developed by Anthony and Coupin (1925–26, 1929).

This time, the topic is one of ontogeny. The object is to measure the brain of the immature in terms of its proportion to that of a definitive adult.

Let the adult brain weight $PE = kPS^r$, where $PE =$ "*poids encéphalique,*" $PS =$ "*poids somatique,*" $r = .25$. Let PS' be body weight at some immature stage. Calculate $k\,PS'^r = PE'$. Let PE' and PS' be known empirically. Then, $\frac{PE}{PE'} = \frac{k\,PS^r}{k\,PS'^r}$, which fraction is the "Index of Cerebral Value." In a 5-months-old fetus, it is far < 1; at about age 7 years, it is a maximum (> 1); by age 30 years, it drops back to 1.

If, on FIGURE 1 or 2, we were to draw a straight line of slope .25 (*i.e.,* 14°–15°) through a mean adult value, the reason for this rise and fall would be clear: such a line would cut the curve of FIGURES 1 or 2 twice.

In all justice to these authors, they admit that their device is a "stratagem." At the same time, it seems to me to carry the interpretation of figures too far when the fact that, after age 30, their index falls more in men than in women, and is coupled with their observation, "*ce-que l'on sait* (italics mine): *la femme semble conserver mieux que l'homme, pendant le période de décrépitude, l'intégrité de ses fonctions intellectuelles.*" By their own figures, the Index reaches the highest point in 7-year-old girls. Hence, it would seem to follow that the

7-year-old girl should be the culmination of man's intellectual development.

It must occur to the reader that the occasional use of $r = .25$ for one kind of comparison within the human species and that of $r = .56$ for another, even though the authors are not the same, leaves us wondering about the logic in the whole system It seems opportunistic, when choice of one or the other depends ahead of time upon the results one desires. Harris (1929) has experimented with the device of Anthony and Coupin; but whether the device has enough substance to be a safe foundation for anything, may be left to the judgment of the reader.

e. The fifth criticism applies to the developments of the Dubois system by our Dutch contemporaries.

I have reproduced Brummelkamp's data for ungulates and rodents (TABLES 13, 14), both to make the present criticism clearer and also because the data will be useful later on.

FIGURES 13 and 14 show the scale made up of a set of lines sloping at $r = .56$, with the vertical distances between them equal, as already explained. The rise in cephalization thereby is indicated as saltatory. But is the sloping scale convincing?

(1) The vertical distances are so small that "the farther 'tis from England the nearer 'tis to France" Some of the points—in fact, all of them—are data from very small populations; the probable errors must be large A fundamental of Dubois' exponent is that species within a genus must be equally cephalized. However, here the genera *Mus, Arvicola, Cervus* are indicated as having more than one cephalization. This renders suspect those genera represented by only one species.

(2) What virtue is there in averaging all specimens that apparently cluster about any single diagonal, when they follow no taxonomic scheme? If this can be done within an order, why could it not also be done where two orders overlap? What do you have after doing so?

(3) In fact, had the sloping scale not been superimposed, we should have seen a scatter-field with a curved axis of orientation. This is an important matter, which Brummelkamp has not exploited. It is cardinal to our exposition later on.

f. There are certain further criticisms to be had from von Bonin, who expresses some of the foregoing as well. But, as he follows up his objections with a constructive analysis of his own, it would be simpler to state his entire position at one time. This we now do.

B. VON BONIN'S CORRELATION COEFFICIENT AND REGRESSION FORMULA FOR THE MAMMALS AS A CLASS

1. Description

Von Bonin (1937) reviews in part the history of the problem of brain/body weight relationship, from Manouvrier and Snell to Dubois; then he says (*Ibid.* p. 380): "While Manouvrier's work shows all the care and critical self-control that are the signs of a great mind, Snell's and Dubois' work had laid itself open to criticism on several points First, we have no right to introduce such conceptions as psychencephalon and somatencephalon as measurable quantities. Secondly, it is tacitly assumed that higher intelligence is due to (or associated with) greater brain weight or greater weight of the psychencephalon, and the cephalization coefficient is then taken as a measure of intelligence. But nobody has ever satisfactorily defined intelligence in such a way as to include animals, and, moreover, we know the behavior of only a very few animals with some degree of accuracy. To suppose, therefore, as Dubois does, that each one of certain pairs of mammals has the same intelligence as the other one, is merely begging the question. Clearly, we shall have to avoid all references to intelligence and all preconceived hypotheses about the relations of mind to the brain and define our task thus: to find out whether there is any law regarding the relation between body-weight and brain-weight expressible in numerical terms."

Von Bonin considers that the piecemeal pairing of animals, in Dubois' manner, is not the right approach. Instead, he tackles the problem immediately at the class level. He sees the mammals first in the aggregate; a chart of all species (after the fashion of our FIGURE 21) is to him a correlation scatter-diagram. He asks whether there is a constant correlation over the entire range of the mammals, which can be expressed by a single coefficient. This would demand a straight-line axis. Von Bonin tests the rectilinearity thus: The correlation ratio $\eta = .9449$. By Blakeman's test, $\xi = .196 \pm .053$; and by R. A. Fisher's, $N(\eta^2 - r^2) = 22.56$ for 18 areas, "which leads to $P > .10$ for the probability of the correlation being linear." Thus, he calculates the correlation coefficient: $r = +.83459$, which is high, though not perfect, as one would expect. Hence, his regression formula for $E = KS^r$:

$$\log E = .655 \log S - .75 \pm .312/N,$$
$$E = .18\, S^{.655}$$

(where S = body weight, E = brain weight). This exponent, .655, is nearer to Snell's of 2/3 (= .666) than it is to that of Dubois.

He concludes (p. 388): "Former attempts to analyze the relation between body weight and brain weight suffer from 3 deficits: (1) they presuppose a correlation between intelligence and brain-weight, (2) they make suppositions about the intelligence of animals which are unproven, and (3) they are based on a conception of cortical functions which can no longer be considered valid. The attempt has here been made to work out the correlation between brain-and-body weight, using the same formula ($E = kS^r$) of relative growth as former authors, but taking into account the complete mass of data at present available. There is a close correlation between the logarithms of brain-and-body weight, and this correlation is linear. Brain weight increases as the 0.655th power of body weight The value of the cephalization coefficient K differs from species to species Whether or not this is an indication of the intelligence of animals must be left to the psychologist to answer."

Von Bonin does not examine piecemeal any taxonomic divisions within the class Mammalia. Hence, such schemes as those of Brummelkamp (see our FIGURES 13, 14) have no counterpa. . He eliminates the possibility of any evolutionary interpretation, in the sense of Brummelkamp, or from the standpoint that, within a given order, the smaller-bodied animals with the lower cephalization coefficients can reflect a stage in weight-proportions once passed through by the larger-bodied, more highly "cephalized." Through the entire mammalian constellation of FIGURE 21, he passes a straight-line axis of slope .66. Then, for any given animal species, $y = ax^{.66}$. If x and y be known, a can be calculated; if x and a are known, y can be calculated. Von Bonin calculates a for 115 species of animals. In this situation, a becomes merely an auxiliary for ascertaining y from x; but it is unfettered with any of the rationale which the Dubois school attempted to lay upon it. Thus, von Bonin is antipodal to the Dubois school in his use of modern statistical technique; also in his assumptions, and the point from which to launch an attack upon the problem. His objections to Dubois, I think, are well taken. On the other hand, whether his remedy is equally good, is a matter worth examining.

2. Criticism

a. That Dubois was too "piecemeal," must be true; but is not von Bonin perhaps too "wholesale"?

While we have no device for measuring, for instance, the intelligences

of gorillas and lions (they are approximately equal in body size), I, for one, am reluctant to believe that the larger size of gorilla brain means nothing at all in the matter of relative intelligence. One does not have to accept the Brummelkamp scale of $a \log \sqrt{2}$ before being aware that, on a field such as those of our FIGURES 15, 16, 18, 20, animals of comparable body size do seem to follow some rational scheme of brain size. Thus, edentates are lowly as compared with carnivores; so are marsupials; Primates are very high; Reptilia and Amphibia are exceedingly low.

If we are content to draw an axis of slope .66 through the entire mammalian constellation, we cannot avoid wondering about a certain parity of cephalization coefficient among all forms, at a given distance above it or at one below it. That would throw together some very small and lowly animals with some very large and specialized ones. This, of course, would not vitiate von Bonin's system; but it certainly would give the whole problem of relative brain weights a different orientation. What is this orientation?

b. A second deterring thought comes from the first. Very shortly, we shall offer a system of curved lines that pass through the mammalian constellation, but which, in the greater part of the constellation, will be very gradual. Now, if this possibility be granted, pending its demonstration, then we may say that von Bonin's test for rectilinearity gave a spurious result. That test simply reported that the *chances favored* a straight-line axis over a curved one; but it did not eliminate—and von Bonin does not claim it did—the lesser chance. In the present instance, we shall soon attempt to show that the "lesser chance" obtains.

c. Von Bonin, in practice, disregards the potential testimony of each taxonomic subdivision in the constellation. I have termed this a "wholesale" method of solution. FIGURE 21, which reproduces his tables and agrees, therefore, with his own illustration (*ibid.*), is composed of the following entries: 40 primate species (23 genera); 9 prosimian species (5 genera); 3 insectivoran genera-species; 4 chiropteran species (3 genera); 18 carnivore species (12 genera); 17 rodent species (15 genera); 12 ungulate species (10 genera); 8 cetacean individuals (7 species in 6 genera); 4 edentate species-genera; 4 marsupial species (3 genera).

This is certainly no balanced ration. If the mammalian constellation actually contains trends such as that proposed by Brummelkamp, or that to be developed shortly, this representation would conceal it.

d. Moreover, in the figure by which von Bonin illustrates the mammals, he omits the line that should graph his formula. Therefore, in FIGURE 21, I have added it to the plot. Obviously, it does not fit the data in the lower-left end of the constellation. Had the data been such as to secure equitable representation to all mammalian orders, this would have been even more noticeable. The only way that it could be achieved would be by weighting the data. This von Bonin did not do.

What I consider to be the shortcomings of von Bonin's solution, and those of the solution developed by the Dubois school, will be understood better if we now turn to an alternative—and this the more, because first, it developed out of a growing dissatisfaction with the Dubois view. Then, and only after it had been formulated out of all the raw data I could find in the literature, von Bonin's study came to hand. That study challenged to a careful review.

We may now turn to an alternative hypothesis. From its standpoint, we shall also try to reinterpret some of the phenomena indicated by Dubois and his followers; after that, we shall reconsider the data of von Bonin.

C. THE EXPONENT OF CEPHALIZATION

1. The Exponent as a Second-Degree Parabola

Dubois operated by pairing genera or species of closely related animals of unequal body size, but assumedly about equally intelligent. Then, by way of algebra, he did the equivalent of joining their logarithmic dead centers (on a log-log grid, their points $P{:}X, Y$) by straight lines, and measured the slope. Then, believing that these slopes were sufficiently close to warrant it, he averaged them all and obtained approximately $r = .56$.

Von Bonin plotted a large number of points representing (unevenly, as already shown) a comprehensive miscellany of all extant mammals, tested the entire constellation for a rectilinear axis, found the chances favored rectilinearity, and so calculated a single slope for the single mass.

The method in the present study, developed before von Bonin's paper was encountered, proves to lie between that of Dubois and that of von Bonin.

If we plot on a log-log grid an order, a superfamily, or a family (much as Brummelkamp did, but omitting his $a \log \sqrt{2}$ scale), we have a simple scatter-field with some interesting properties. See FIGURES 13, 14, but more especially FIGURES 16–19.

a. Repeatedly, the larger members of a phylum—bears, tigers, lions, elk, etc.—will fall *below* a slope of .56, if a line with that slope be first passed through the smaller members of the same genus or family.

b. It is possible to trace an axis through each of these constellations abstracted from the mammalian whole, and the axis is manifestly curved and concave downward. At the lower-left end of each constellation, the slope is very steep. Hence, Brummelkamp found occasion to mount his scale for the cephalization coefficient. He had assumed slopes of .56 over most of his fields, so that the lower-left end had to be accounted for by way of saltatory increases in the coefficient. However, he never tried to explain the reverse phenomenon at the upper-right end: the falling-off in cephalization level there, once parallels of .56 have been superimposed upon the field. On the other hand, there is no way of explaining any curvature whatever, once von Bonin's straight line be granted.

c. If the data now be reassembled, as in FIGURE 20, we seem to find a *sheaf* of curved lines, all concave downward, and the steeper the elevation of the line, the more sharply curved it is. (See also FIGURE 22.) Immediately, we realize that no scheme of .56 slopes could be justified in the primate portion of the total constellation.

Thereby, a family of parabolas $Y = A + bX - cX^2$ suggests itself, in which the relations between b and c determine the slopes of each and all. The $-cX^2$ becomes a correction-term to the straight line $Y = A + bX$ (whether of the Dubois or of the von Bonin system) to take care of the downward concavity: the defection of the larger animals at the upper-right, and the steep ascent among the smaller animals at the lower-left.

2. Other Forms Tested

Further considerations, I think, will strengthen the choice of the parabola. Meanwhile, it is well to note briefly that other types of curve have been tested, but found less satisfactory:

$$y = ax^b, \qquad y = ax^b + c, \qquad x = ae^{by} \qquad \text{or}$$

$$y = \frac{\log x - \log a}{.4343b}, \qquad \text{and} \qquad y = a + bx + c \log x.$$

(In all these cases, y and x are general terms for the formula types. Naturally, we should substitute Y and X). While this far from exhausts the possibilities, it does narrow the choice.

We turn now to descriptive analyses of the mean trends or axes in the primates, the carnivores, and the artiodactyls.

3. The Primates

a. It is not unreasonable to suppose that man has risen from some ancient catarrhine monkey form, although discussion of the point does not belong here. In qualitative comparative anatomy, the catarrhine monkeys, especially the *Rhesus*, are stock material for showing evidence in the matter. The *Rhesus* is considered by some to be a rather generalized extant form, and therefore furnishes a convenient point of departure. In quantitative comparative anatomy, however, as has been stated already, an appeal from a *Rhesus* to *Homo* is much less safe. For it has not yet been shown that *Rhesus*—specifically, in the brain/body weight relationship—has stood still from the Tertiary day when the *Homo* line began to draw away from some common catarrhine field. The cercopitheques—at least, their less evolved, more generalized forms—quite reasonably may lie closer to the common ancestor than does *Homo;* nevertheless, proximity is not identity. More specifically, we do not know that, when man's true ancestor was the size of a *Rhesus* macaque, his brain was the size of a *Rhesus'* brain

But since we are making a study in quantitative comparative anatomy (quite confessedly, in the hope that it may furnish some groundwork for later, truly phylogenetic studies) we shall find it convenient to appeal to the catarrhine monkeys. Now, they vary widely in brain and in body size. But we need some dead center. After weighing the factors, I have chosen the mean brain and mean body weight of my cercopithecid data, but omitting the Cynocephali because of their obviously high specialization. The choice, therefore, is subjective, and future events must be its judges. Hereinafter, "Cercopithecidae" shall be the (unfortunate) term used to mean these monkeys, with the Cynocephali omitted and treated separately.

By extension and speaking the same way, the primates are descended from insectivores. We have no comprehensive data on these, but Crile and Quiring give the means of 68 shrews, which are very small animals.

There are a few other scattered data of insectivores; but in a log-log plot they would not fall far away from the shrews. In the present experiment, let the shrews suffice.

The point for man is from the pooled adults (sexes combined) of the data of Crile and Quiring. Their list includes individuals of various "races."

The three points—*Homo*, Cercopithecidae, *Blarina*—are joined by a curve, concave downward, expressible as a parabola having the formula

$$Y = -2.2417 + 1.549X - .0899X^2$$

or

$$y = .005732x^{1.549 - .0899 \log x}.$$

While any three points can be connected by a parabola, it is immediately striking that the marmosets, the Cebidae, and the baboons—three discrete points—range themselves very close to the curve. See FIGURE 15 and TABLE 19.

The straight line $Y = A + bX$ is but the central case of

$$Y = A + bX \pm cX^2;$$

for there, $c = 0$. The antilogarithmic equivalent of $Y = A + bX$ does not have to be written $y = ax^b$; it may also be written $y = \text{antilog}_{10}(A + bX)$, or $\varepsilon^{2.303(A+bX)}$ Similarly, and for convenience, let us write $y = \text{antilog}_{10}(A + bX - cX^2)$, or $\varepsilon^{2.303(A+bX-cX^2)}$, and call

$$A + bX - cX^2$$

the exponent of cephalization.

b. As striking a feature as any in the primate line is the spacings between the several stages of animal types. They are listed in TABLE 20. The ratios between man and baboon and between man and Cercopithecidae are the *lowest* in the list: *i.e., the greatest proportionate rises of log brain to log body weight occur anywhere but between man and the forms next below him.* (Incidentally, these ratios correspond to the formula

$$\frac{\log E - \log e}{\log P - \log p} = r.$$

In the case of baboons/Cercopithecidae, one might have presupposed "equal cephalization" and hence have expected $r = .56$, provided the Dubois scheme be correct.) This progressive loss of slope is a crude indication, if such be needed, that the several morphological levels are joined by a curve concave downward, whatever be the actual type of formula that fits it best.

At least one thing is certain: there is nothing "aberrant" or "exceptional" about man's cephalization, unless it be the fact that he occupies one extreme of the mammalian front. By such definition, the extreme would always equate with aberrant, no matter what the animal might be.

c. Granting the *Homo* line of mean tendency, how broad a band on both sides of that line would be tolerable for the variational proportions of man and of his sub-human lineal ancestors?

In FIGURES 15 and 20 is an enclosure or a line indicating a tolerable range in the brain-to-body logarithmic ratio for *Pithecanthropus erectus*. (Brain weight has had to be estimated from cranial capacity. For a discussion, see Martin (1928), II: 743f. There, these figures occur for modern man:

	Brain weight ratio of gm/cc	Cranial capacity
Welcker 1886	.91 at	1200–1300 cc
	.95 at	1600–1700 cc
Bolk	.737–.94 at	aet. 30 years
Manouvrier	.87	general average

Experiment with figures in the literature has led me to adopt .876 as a workable figure).

Some data on Palaeanthropoi are assembled in TABLE 18. Unless otherwise indicated, the brain weight estimates are my own. The cranial capacities are taken from Hrdlicka (1930), except for Weinert's, which is from his *Ursprung der Menschheit*.

The reader may shift about these "brain weights" on the charts *ad lib.*, to try them with different putative body weights. Now, if increase of brain weight demands a certain increase of body weight and increase of brain complexity demands a certain increase of brain weight (This seems to be a cardinal implication of our chart, so that a creature of gibbon size, yet with the brain complexity of a man, seems ruled out as a normal thing, despite the occasional occurrence of Tom Thumbs), then *Pithecanthropus erectus* cannot be ancestral to extant man. His body is too large for his brain, just as, to a more exaggerated degree, that of an ape is too large for his brain to represent or reflect the subhuman proportions of any ancestor of man. In terms of the graph, it seems too much to demand that the *Homo* line pass first through *Pithecanthropus*, then suddenly take a geniculate turn vertically upward. As a matter of speculative theory, when a body develops ontogenetically, it budgets a certain proportion of its total growth-energy toward developing the size and complexity of its central system of control. The present study supports the notion that the budgetary balance between the total growth-energy of the soma and that portion appropriated to brain is too fundamental in the biological economy of the organism to allow the total body to remain phyletically stationary in size, while the brain growth budget is revised drastically upward. Rather, the alteration of the brain-to-body ratio is conceived as being impossible without concomitant alteration of absolute size in both body and brain. This is the obverse and reverse of a common biological fundamental. When man's direct ancestor had a brain the size of that of *Pithecanthropus*, his body was smaller: according to the *Homo* parabola, about 40,000 gm. But what the *complexity* of that brain would have been, as compared with that of *Pithecanthropus*, there is, of course, no telling.*

* See "Addendum" at end of Section IV.

4. The Carnivores

We assume:

(a) The carnivores are derived from small insectivores

(b) Among progressive stocks, the smaller (carnivore) is more likely to be the conservative member of his group: *e.g.*, of the Felidae, with their great body range, the smaller cats will be nearer the ancestral proportions than the larger ones. Likewise, among the Procyonidae, Canidae, Ursidae, the first will be most nearly of ancestral proportions These assumptions may be false, or only inaccurate; nevertheless, the experiment is worth trying.

From Crile and Quiring's adult racoons, weasels, and the same shrews as before, we obtain the parabola:

$$Y = -1.9227 + 1.283X - .0844X^2.$$

See FIGURE 17. The points in the figure are the data of TABLE 15, their numbers being those of the entries in that table None of the specimens beyond 15 (*Procyon lotor*), to say nothing of most of those below it, have had anything to do with formulating the line. Yet I do not believe one could ask for a better fit.

5. The Artiodactyls

FIGURE 18 shows them running approximately the same course as the Carnivora, suggesting that the modern survivors of these two orders (both of which have creodont ancestry) have been running neck-to-neck in an evolutionary race: predator and prey, balanced thus far, so that the former has not exterminated the latter. However, the Tragulina are placed a little too far to the right, just as the lemurs are, to fit a parabola of mean tendency. The gap between the shrews and the smallest antelopes is too great for a primary case in demonstration. Now, however, with the Carnivora so amenable, an essay on the artiodactyls is fairly safe. The Antelopinae are represented by a great body range. So, taking the shrews, *Cephalophus,* and the three genera 8, 9, 12 (TABLE 13), which are clustered, we obtain:

$$Y = -1.9284 + 1.289X - .0838X^2$$

and the curve is in FIGURE 19. From TABLE 13, the other points are added as checks. The line is practically identical with the carnivore.

6. A Criticism of the Curvilinear Cephalization Exponent Based upon the *Homo* Line

Consider FIGURES 15, 17, 19, 20.

a. It is as important to understand what the *Homo* parabola is not, as to understand what it is.

It does not represent an *average* trend among primates, but rather some kind of higher extreme; a line by which man can be reached. Had we used Hrdlicka's (1925) orangs, instead of Crile and Quiring's humans, we would have had, out of otherwise the same data:

$$Y = -2.6925 + 2.0518X - .2007X^2.$$

(The orang values of $x = 54507$, $y = 335.7$ were obtained from the unweighted average between means for 8 male and 12 female adults.) On the other hand, the lemurs occupy, far down the scale, a position relative to the *Homo* parabola analogous to that held by the orangs, far up on the scale, relative to the same *Homo* parabola, but we would not be justified in trying to pass a lemur parabola through the Hapalidae, as we have done with the orang and *Homo* lines.*

The scatter-field of adult Cercopithecidae in FIGURE 15 shows what tremendous latitude in brain/body weight ratios exists in the primate constellation; or, what a wide range in the constants of *b* and *c* should exist among primates. This is a problem in variance. When coupled with analogous data in other orders, it carries this study into more advanced levels which cannot now be explored. But let us state carefully that, in spite of all the parabolic treatment in this study, we must *not* imagine the differential evolution within the primate stock as following a set of discrete parabolas, so as to make man ascend strictly by way of one trajectory while the orang has followed another close by. We cannot be such purists as to imagine that to be the case. Rather, what we are attempting to develop is the picture of a *band* in the primate portion of a total mammalian constellation, having parabolic curvature, and a breadth as yet undetermined at any point along its course. That band might be visualized as the limiting outline of a whole sheaf or spray of parabolas having various inclinations and curvatures. But even that would be a fictive device for trying to grasp the biological entity behind it. That entity is one of range in variation of body-to-brain weight throughout the entire period when primate genera and species are gradually evolving and differentiating. The mechanism is, of course, that of unceasing matings and breedings of countless individuals. The proportions peculiar to man or peculiar to orang, etc., finally emerge, but even they have a range tolerance in the matter of brain and body weight. The *Homo* line is an *axis of mean trend for attaining man from out of the insectivores.* The axis is a line on which, *for purposes of calculation,* we may consider all entry points to have been clustered. It is an "*as if*" line. At any level with-

* FIGURE 15 amply demonstrates why it was but an accident that Dubois obtained a slope of .56 in his Primates, when he paired the orang and the gibbon.

in the primate band which is below both the definitive human and the ape and, therefore, by right of title common to both, if we had the requisite data on their ancestors, those ancestors probably would be all mixed together, and identifiable at some differentiating epoch only by the qualitative taxonomic methods of the paleontologist. Quite conceivably, however, if we did have all these ancestral entries, when we stand at the spot where the population is about to differentiate the humanoid and the anthropoid, those destined to be progenitors of man might tend to be more frequent on the upper side of the primate band, while those destined to produce apes might trend towards concentration on the lower side. Whether this speculation be justifiable, probably must remain unresolved for a long time.

b. Necessarily, in the carnivores and artiodactyls, we should have to explain the positions of such forms as Tragulina and Mustelidae in the same way that we have handled the Lemuroidea, the Anthropoidea, and the Cercopithecidae. The Cercopithecidae, Mustelidae, and even the Lemuroidea have their own local axes lying athwart the parabola of the cephalization exponent. They recall Lapicque's dogs and his ontogenetic exponent in its position relative to the phylogenetic slope of .56.

7. A Line of Mean Tendency in Modern Mammals

None of the other orders, not even the rodents, which are well represented in point of genera, but which present certain peculiar difficulties, yields as ready returns as the three foregoing orders Yet parabolas can be fitted to them that are consistent with those three. Later on, we shall do this (TABLE 26), but such testimony cannot be used to establish the case of the parabola.

On the other hand, the entire mammalian constellation does lend itself to calculating the mean trend of *extant* mammals. The method here followed has attempted to weight the various genera equally. After a formula has been calculated from my own collection of data from the literature, the experiment has been repeated on von Bonin's collection.

Consult now TABLES 24 and 25.

The theory underlying the method is simply this. Each animal kind is considered as representing a situation. Thus, no matter what its success or failure in breeding large numbers, or even in differentiating species out of a genus (surely the brain-body relationship could not differ palpably among species of one genus), the brain/body weight

relationship is a biological reality. Now, Dubois and his followers used mostly the genus as a unit of comparison. It should be safe to make all entries in terms of genera. Nevertheless, we really should have the same number of genera in each order, if it is orders that we wish to formulate So I have applied this formula:

1. Take the *mean* of a *genus* as the unit of entry.

2. Count the number of genera in each order Let the number for the most numerously represented order be standard

3. Divide this number by the number of genera in each of the orders, to obtain a multiplier or operator: *i e*, let C be the greatest number of genera in the best-represented order, and N be the number of genera in any one of the others. Then,

$$\frac{C}{N} = W,$$

and W is the operator by which to weight each Y and X in the order. Now, a genus occurs in a certain spot on the graph. The orders are more densely represented in one part of the field than in another. We wish an even distribution. The range of X is therefore divided, at intervals of .5, into 11 classes, and all the genera within each area so defined, no matter what their orders, are grouped for an average. For instance:

GROUP 3

Genus No	W	WX	WY
12	1 304	3 020	1 180
13	1 304	3 060	1 100
23	1 304	3 240	1 270
40	1 304	2 870	924
74	1	2 484	774
75	1	2 201	613
90	1	2 301	201
96	1	2 033	452
N = 8	Σ = 9 216	Σ = 21 209	Σ = 6 574

$$\frac{\Sigma WX}{\Sigma W} = 2.304 \qquad \frac{\Sigma WY}{\Sigma W} = .714.$$

These two quotients stand for the "dead center" of Group 3. A parabolic formula can then be calculated from the 11 groups.

In making my revised calculation, after reading von Bonin's paper, I added his insectivore genera to mine; also, I excluded *Balaenopterus*, as he had done, but saw no reason for excluding *Homo*. Furthermore, since the species of *Felis* cover so wide a range, and von Bonin lists them as several genera, I grouped them into four body-sizes and let

these stand for genera. Von Bonin lists *Asinus asinus* as a genus, while I left it *Equus asinus*. In handling my own data, I could not bring myself to make a separate order of the Prosimiidae; although, in calculating von Bonin's collection, I left them separate to conform with von Bonin. My own collection contains at least 116 units (tantamount to genera), namely:

23 Primates, 23 Carnivora, 22 Ungulata, 30 Rodentia, 3 Pinnipedia, 2 Cetacea, 4 Edentata, 4 Chiroptera, 5 Insectivora.

Then $C = 30$, and each of these numbers equals N, whence in each we obtain $\frac{C}{N} = W$. At last, the formula is:

$$r = -1.808 + 1.154X - .0543X^2.$$

See FIGURE 20. It seems a very fair fit.

Repeating on the 82 units or genera of von Bonin's list less the giant Cetacca and less the 3 marsupials, which are not Eutheria, we obtain, from 79 units or genera:

$$Y = -1.4045 + .8724X - .01363X^2.$$

See FIGURE 21.

The data of von Bonin give a flatter parabola, but the first passes closer to the Balaenoptera, which were included in the calculation. The first also converges better with the ordinal lines at the lower-left end.*

We have observed previously that statistical tests for the rectilinearity or curvilinearity of a constellation (such as the mammalian of FIGURE 20 or of FIGURE 21) simply measure a *probability* of the condition being the one or the other. These tests do not *eliminate* the alternative possibility. But before we can penetrate further into the problems actually confronting us, we are in duty bound to test our own mammalian constellation by the conventional methods.

In this matter, we shall have to give to the genera of each order their weightings as per TABLE 24, then calculate the correlation coefficient r and the correlation ratio η and use one of Blakeman's tests for a significant difference. We obtain:

$$Y = +.966 \qquad \eta = .969.$$

Applying now the formula,

$$\frac{d}{\sigma} = \frac{\eta^2 - r^2}{2\sqrt{\frac{(\eta^2 - r^2)[(1 - \eta^2)^2 - (1 - r^2)^2 + 1]}{N}}}$$

* In FIGURES 20 and 21, there appears to be too heavy a field below the curves. But marsupials and extinct mammals have been entered in the field without being included in the formulation. A further distortion, not measurable to the eye, comes, of course, from the use of the frequencies W of TABLE 24.

we obtain .624, which is not a significant figure. Therefore, if all there be behind the constellation is only a statistical correlation, von Bonin is right: the axis is probably rectilinear.

But since we have approached the analysis by way of orders or stocks, where any line that passes through their samples is manifestly curvilinear, the question is legitimate: Cannot their aggregate permit a *deceptively* straight line to be drawn through the whole?

My own belief is that the rectilinear axis is a deception. Let us recall some descriptive remarks from the discussion of "pseudo-phylogeny" (Section III, Subsection 2, Preliminary a) and picture in somewhat better detail the upsurging of the true "phylogenetic" lines through the pile of transparent sheets representing the past epochs with their now-fossilized forms. In this connection, it is rewarding to consult FIGURE 13, the ungulate scheme reproduced from Brummelkamp, and note the positions of the small, white circles that are entries of extinct forms; also, anticipatorily, FIGURE 22, where *Diplobune* and *Anoplotherium* are again entered. In the scheme we are now imagining, these extinct forms (and countless others analogous to them) would not occur on these surfaces actually before us, but would be buried somewhere in the depths of our pile. Using the artiodactyls as an example, the indication is that, if we had all the curves on all the sheets of our pile, they would generate a curved *surface*, standing on edge, so to speak, and cutting upward through the pile of sheets: a surface located in the third, or z-, dimension. Moreover, the now-extinct forms seem to describe curves in which the brain weight is less, relative to body weight, than it is in extant forms. In other words, as we come up through our pile, the curves also tend successively to become more elevated with respect to the y-axis. Modern ungulates have a higher brain/body weight ratio than extinct ones.

Now, with this element of behavior in mind, consider the diagram of FIGURE 22 as a whole, and consider it as the top sheet of the pile Then, in the lower-left region, there still exist mammals of comparatively low cerebral organization, sloths, kangaroos, *et al.*, which are also relatively small in body size. Peering down through the pile, however, we perceive giant sloths, glyptodonts, kangaroos, wombats, and many others, such as *Baluchitherium*, *Uintatherium*, the last of the titanotheres; to say nothing of Irish elk, the Imperial elephant, *Megaladapis*, and *Gigantopithecus*. It is evident, then, that, unless a phylum produces members of ever higher proportion of brain weight, its only representatives that survive eventually are animals of relatively small size; otherwise, the entire phylum becomes extinct. If, then, we

had all the lines of all the orders from at least early Tertiary to Recent, they would shape for us a block as though it had been carved. It would be curved in a way hard to describe. It would be standing on end, and leaning at an angle. As we look down upon it, through our pile of transparent sheets, we get the impression of a "cliff" facing down and to the right: a face where Nature has eroded away the surface that would have been formed by the aggregate of the right ends of the lines, in that region, the lower-right, where large body and small brain would occur. For marsupials, edentates, rodents, Chiroptera, Hyracoidea, are represented at long last, on top, only by small forms, while Multituberculata, Tillodontia, Amblypoda, Notoungulata, *et al* are completely extinct. The most highly cephalized line, that of the Primates, which belongs to that surface of our model that faces *up-and-left,* does not contain, and never has contained, the largest of all mammals in point of total body weight.

Add now the complication that each of the lines, and the surfaces which we are imagining that they generate by their close contiguity, is but an axis of means. To each side of each line should lie an accompanying province for everything that deviates from the mean.

In short, had we carved a block of wood to represent the history of cephalization in mammals, its scheme of construction might be made up completely of a system of curvatures; then, as we whittled or sheared off a large chip representing what the extinct lines would have become had they been allowed to continue, we might well have ended up with a certain surface on our top sheet, the shape of which would permit us to calculate a certain long axis for it. This straight-line axis then should be, I think, the regression-line calculated by von Bonin.

Meanwhile, by confining ourselves to the much-trimmed plant that still survives, we can obtain, I think, some idea of the range in log brain weight that exists today with respect to the range in log body weight. Using as arrays of Y the eleven groups we have formed by dividing the scale of X at intervals of .5, we can obtain a σ for the mean of each array (M_Y). The regression of σ_Y on X is:

$\sigma_Y = .2072 + .01756X$ (unweighted by the operators W);

or $\sigma_Y = .2547 + .00539X$ (weighted as in $\Sigma f \cdot \sigma_Y = \Sigma f \cdot a + \Sigma f X \cdot b$, where $f = W$ properly applied to each genus.)

Thus, the standard deviation increases practically not at all as X increases. If, then, σ_Y of each array be divided by the M_Y for that array, the quotient CV_Y follows a very regular negative parabola:

$$CV_Y = 1.699 - 1.3135M_Y + .2545M^2_Y,$$

so that the range of variation in Y narrows rapidly with increase of Y and X in the extant mammalian constellation. (At $M_X = 2.304$ and $M_Y = .714$, $CV_Y = .374$; at $M_X = 6.23$ and $M_Y = 3.2035$, $CV_Y = .0903$.) This is not unexpected. For, as FIGURE 20 shows, with rise in X fewer and fewer orders participate: as we have just been saying, they never produced giant representatives, or else their giant members have become extinct. Nature has not been indefinitely patient with big bodies plus small brains; nor, for that matter, has she ever produced superlative brains in superlative body sizes. Neither has she produced superlative brains in tiny bodies.

D. A STATISTICAL ANATOMY OF THE MAMMALIAN CLASS, ASSUMING THE VALIDITY OF THE EXPONENT OF CEPHALIZATION

INTRODUCTION

1. The adoption of $Y = A + bX \pm cX^2$ as the type-formula for representing the relationship between brain and body weight is, then, frankly empirical only. Whether it can be rationalized theoretically, I do not know. At least, as will develop soon, it yields certain further results that are consistent. This much, I think, is already reasonably secure: the axis of mean tendency in order or class is a curve, concave downward, and therefore changing slope at every point.

From now on, we shall behave *as if* the second-degree parabolic formula were the most acceptable type and explore the consequences.

2. The formula $Y = A + bX - cX^2$ is analogous to that of a shell trajectory, which is often written $T = s_0 + v_0 t - \frac{1}{2}t^2$, where T is the trajectory, s_0 is the initial distance or position, v_0 the muzzle velocity, t the time, and g the gravitational force. But in the exponent of cephalization, discrete factors cannot be so isolated and labeled. On the other hand, if the Mammalia originated from reptiles, their parabolic exponents, which (in FIGURES 20, 22) spray outward and upward and to the right, presumably must have had an "initial distance." Later on, we shall try to render this particular a little more precise. For the present, we shall study the relations between the terms containing X.

The correction term cX^2 progressively subtracts from what otherwise would be a rectilinear slope. Hence, *whatever be its biological antecedents, if it has any,* it is the analogue of $\frac{1}{2}gt^2$, and $c = \frac{1}{2}g$, or $2c = g$.

The velocity of the exponent of cephalization is, of course, given by:

$$\frac{dY}{dX} = b - 2cX.$$

The value will be greatest for the smallest of the mammals. In fact, as we shall see later, the closer we come to a reptilian level in the general region of which the mammals presumably departed, the greater is this velocity.

The velocity, at any given value of X, is highest in the primate parabola. It departs farthest from the mean of mammals on the plus-side. This adumbrates that aspect of the problem, again, which deals with the variance, but we shall postpone that aspect. For the present, an indication of the comparative velocities of the various parabolas can be had by finding the point at which the rate of increase in log brain weight has fallen to equality with that in log body weight; that is, where the slope of the parabola = 1, or forms an angle of 45° with the abscissa. Let us call that the point of *isauxon*. Then, when $b - 2cX = 1$, X equals:

in the *Homo* line,	3.05
in the carnivore line,	1.63
in the antelope line,	1.72
in the mean mammalian line,	1.42.

The primate isauxon point occurs, then, comparatively late, somewhere between the hapalid mean ($X = 2.44$) and the cercopithecid mean ($X = 3.30$). The shrews occur at $X = 1.2396$, before these isauxon points, and therefore presumably in a region where, when the early primitive mammals existed, the log brain weight was still increasing faster than log body weight. The deceleration $-2cX^2$ had not yet brought the velocity down to 1. This comments on the original and elementary phenomenon that was partly responsible for the investigations of Manouvrier, Snell, Dubois, and the others of their day: the brain weight/body weight ratio of some Insectivora is greater even than that of man.* But the present explanation obviously differs from the solution of Dubois. And the mammalian scheme, a falling-off in the ratio as animal bodies increase in absolute size, will later on be found to contrast with the reptilian scheme.

At the same time, FIGURE 20 demonstrates that the most highly evolved of the mammals tend to occur even more to the right and higher up. *Cerebral organization, thereby, seems to demand an absolute increase in body size to support it.* Yet this increasing cerebral complexity goes with a falling ratio of log brain weight to log body weight, so that even giant man rises no higher in brain weight per body weight than his putative insectivore ancestor.

* More precisely and graphically, if a slope of 45° be passed through the Hapalidae, even the primate slope loses height rapidly enough to place man below that line. In marmosets, the ratio of brain weight/body weight is .039; in man, it is .021

This is but a statistical comment to some very profound and obscure neurological processes among the mammals.

Almost as striking is the further fact that the body-size range of 1000–10000 gm. includes the bulk of extant mammalian types and an extremely wide range of cephalization. It appears as a sort of focus. Beyond it, to the right, is a region once held by mammals now extinct, and above that region one which has been "settled" in latter ages by the more specialized, large-bodied, extant forms. In this region, the large artiodactyls and carnivores now appear as close to the bottom of cephalization for mammals of that magnitude, because the archaic giants are extinct. This partial erasure of the mammals on FIGURE 20 is but a reflection of the psychic-armament race which the class has been conducting intramurally.

E. THE REPTILIAN SUBSTRATUM

1. A Reptilian Cephalization Exponent

FIGURE 20 suggests that, with fuller data, we might reasonably expect other ordinal parabolas after the fashion of those just presented. A little later, we shall attempt what is frankly an approximative reconstruction of several. Before that, it is desirable to relate such parabolas as we already have to a possible reptilian line.

For this experiment, we have the data which Dubois and others used in obtaining the phylogenetic exponent and the cephalization coefficients of reptiles (and of amphibians). (See TABLES 21, 22.) These are extant reptiles, just as the mammals we are using are extant.

It will be noticed (FIGURE 20) that, in their brain/body weight proportions, the smaller reptiles are identical with Amphibia. In spite of the higher morphology of the reptile, the size of its brain is not sufficiently heavier than that of an amphibian of equal body size to register a significant difference on the graph.*

It will be noticed, further, from TABLE 21 and FIGURE 20, that the reptiles localize rather definitely by their classes, so that, even if we were to follow Dubois' system, the mean cephalization coefficients would tend to separate them as Crocodilia, Lacertilia, Chelonia, Ophidia, in that order. None the less, the range of the individuals within the one order Crocodilia is relatively large.†

* Brummelkamp gathered his Amphibia (they are all Anura) into three groups, according to mean values of k: .01799, .01317, .009391. (These give logarithms which, when divided by $\log \sqrt{2}$, yield $\alpha = -13.46, -12.5, -11.6$; which, to be sure, differ by approximately 1 whole unit.) The reptilian mean values of k range, as TABLE 21 shows, between .0222 and .00927.

† In terms of k, it is 1.659; or, on Brummelkamp's scale, in terms of $\log k$ it is .2198—*i.e.*, 1.460 $\log \sqrt{2}$; *i.e.*, α is nearly 1½. This is more than 1 and less than 2 units of α. I do not know how Brummelkamp would explain such a high range of saltation.

We compute a parabola from extant reptiles, for comparison with the mean parabola of extant mammals. We obtain, from all the data of TABLE 21,

$$Y = -1.7095 + .3679X + .03613X^2.$$

This is concave *upward:* the log brain weight *accelerates* as body size increases.

Under the Dubois system, this same phenomenon would have to be interpreted as a rise in cephalization coefficient, so that the Crocodilia would stand highest and Ophidia lowest. (Note, however, that the data do not include large Chelonia or Ophidia, such as the great sea tortoises, pythons, etc.)

For further comment on the positive curvature to the reptilian parabola, see Section III, F 2 a, Addendum.

2. Intersection of Mammalian and Reptilian Exponents

FIGURES 20, 24, *et al.*, demonstrate the fact that the mammalian parabolas intersect the reptilian in a region of very tiny reptiles.*

The lower intersection of the two equations:

$$Y = -1.7095 + .3679X + .03613X^2$$
$$Y = -1.8080 + 1.1540X - .0543X^2$$

is $Y = -1.6624$, $X = .1267$. Allowing sufficient latitude to make this point the indication of some limited region, we may still conclude that the mammals must have diverged from some tiny-bodied reptilian stock. The origin of the impulse to devote a larger percentage of growth-energy to brain development must, therefore, have been possible only in bodies of small size. That plasticity is not to be looked for after reptilian bodies have increased beyond some rather modest point. But just where that point was, cannot be said.

Pending, then, any modifying testimony from fossil forms, the data of extant mammals and reptiles indicate that:

Relative brain weight *increases* in reptiles as body size increases;

The animals that eventuated in mammals were small creatures that initiated a markedly accelerated growth in weight of certain parts of the brain;

The acceleration was related to increase of body weight in such a way that increased complexity, particularly of certain parts practically absent in the reptiles, thereby became possible; or, stated conversely, without increase of body size, there could have been no increase of complexity such as has taken place;

* Indeed, it is possible that, had we the requisite data from fossil forms, including the theromorphs, the divergence of the mammals from the reptiles by way of a sprouting neocortex would give the mammalian curve an S-shape with a very short lower limb. The parabolas we are tracing would then be merely the very long upper limbs from an equation higher than the second degree. This, of course, would not vitiate the present development; it would but correct and amplify it.

The rise in log brain weight with respect to log body weight was most rapid when and where the mammalian stock was early, small, and primitive;

Yet the increase in mammalian brain weight has been moderated by a steady subtraction from the logarithm of brain weight expressible as $-cX^2$, where $X =$ log body weight.

When compared with the reptilian behavior, the mammals have taken a peculiar shortcut to large size of brain, which, because it concentrates on certain areas very much restricted in the reptiles, has passed from the quantitative to the qualitative.

There will be a little more to say about this later on. At the moment, this mammalian preeminence and the shape of the curves by which it is traced raise a new and interesting problem. Rather, they place an old problem in a new light:

3. Weights of Mammalian Brain Stems and Weights of Reptilian Brains

If, for want of something better, we roughly term the mammalian brain "neëncephalic" and the reptilian (and amphibian) "archencephalic" we might ask whether, if it were possible to isolate the "archencephalic," portions of the mammalian brain, they would prove to have also participated in the mammalian upsurge, or whether they have rather tended to preserve an archaic weight relationship to the body as a whole.

The question, of course, cannot be answered at all accurately. Nevertheless, the following experiment is suggestive.

Data exist on the weights of mammalian brain portions as divided by the knife. While the comparison obviously is not more than partially correct, we may consider the "brain stem" of some mammals, beside what the total weight of brain would be in a reptile of equivalent body weight, the reptilian brain weight being computed from the parabolic formula just obtained.

a. Starting with man, we have data from Marshall (1892) who constructed from Boyd's tables (in pounds and ounces) the mean body weights and mean weights of total encephalon, cerebrum, cerebellum, and pons-plus-medulla. (See TABLE 23, which abstracts from Marshall.) Boyd's people were clearly undersized of body, so that our computed ratios in column VI presumably are a little high. Anthony (1928) states the ratio to be 2 per cent. We shall use the operator .0206, the average of mean male and mean female of TABLE 23.

Applying it to the mean brain weight of Crile and Quiring's humans —1320.15 gm—we obtain 27.22 gm for pons-plus-medulla.

Anthony (1928, p. 92) reproduces Manouvrier's calculations from Broca's Parisians:

Stature	154 ♂♂	44 ♀♀
	1680 mm.	1583 mm
Weight of encephalon	*1361 5 gm.	1201 3 gm
Weight of hemispheres	1191 0	1045 44
Weight of cerebellum	145 2	131 7
Weight of pons	19 51	17 8
Weight of medulla	6 805	6 36

* There is a small mistake somewhere. The parts add up to 1362 5.

The calculated quantity, 27.22, agrees well with the measurements in this table (pons-plus-medulla).

From this table, the ratios of medulla weight to total brain weight are: males, .00500; females, .00529.

Body weights are not given in these data. But now we can appeal to Crile and Quiring's data, thus:

	♂	♀
Brain weight	1350 gm.	1214
Body weight	65544 gm.	49755 gm
Calc. medulla weight:		
00500 x 1350	6 75 gm.	
00529 x 1214		6 425 gm

These medulla weights agree well with Manouvrier's figures.

b. We add the domestic cat The data are abstracted from Latimer (1938):

	52 ♂♂	52 ♀♀
Brain weight	27 56 gm.	26 54 gm
Medulla plus pons	1 86 gm.	1 79 gm.
Brain/body weight	01031	01126

c. Finally, we bring in the albino rat, from Donaldson's (1924) Tables 140 and 144:

Brain weight at 150 days:	1.933	gm.
Body weight when brain weight = 1.933:	259.1	gm.
Brain "stem" weight (encephalon minus cerebrum and cerebellum):	.373	gm.

Let the weights of these portions of mammalian brain be compared with the calculated total brain weights of theoretical reptiles of equivalent size:

	Body weight gm	Log body weight	Part of brain: gm.	Log of brain part	Reptile	
					Brain weight gm.	Log brain weight
Homo	62080	4 79295	27 22[1]	1 4349		
			6 75[2]	8293	7 65	8835
Cat ♂	2668	3 4262	1 86[1]	2695	99 –	– 0255
Rat	259 1	2 4433	373[3]	– 428	254	– 5949

1 Pons-plus-medulla.
2. Medulla alone
3 Encephalon minus cerebrum and cerebellum

It is strikingly clear, after due allowance is made for the mammalian pons—which in man is particularly heavy, for certain known reasons—and also for the fact that we are comparing at times only the mammalian medulla with the entire reptilian brain, that the mammalian "archencephalon," if it could be reconstituted, would be of the same *general order of magnitude* as that of a reptile of equivalent size. That the mammalian medulla equates neurologically with the reptilian, is of course an erroneous assumption. But let the figures speak for themselves as far as they can.

One comparison we cannot make, yet it would be more valuable than this arraying of a primate, a carnivore, and a rodent. It would be a comparison of brain stem weights with body weights in several stages within the same order or on the same trajectory; *e.g.*, *Blarina*, Hapalidae, Cebidae, Cercopithecidae, Hominidae. Would a trajectory through these data be concave upward or downward? How would it compare with a reptilian line? If we had such a curve, we could, at all times, measure the vertical difference between the curve for total encephalon and that for brain stem. One need not subscribe to the postulate of "somatencephalon" and "psychencephalon," to consider that the experiment would be suggestive.

4. Discussion and Criticism of the Reptilian and Mammalian Parabolas

Judged on the basis of extant reptiles only, the path of this class curves ever upward, which would mean an accelerated enlargement of brain weight as body size increases. But have we a right to assume by this token that the modern crocodile once had an ancestor of the same brain and body weight proportions as, say, a modern turtle or snake?

FIGURE 20 shows a cluster of amphibians and small reptiles at the lower-left end of the reptilian constellation. The reptiles there are small lizards and snakes. It can hardly be imagined that the trajectory of the crocodilian line did not pass close to this region, but could not that trajectory have been slightly concave downward as readily as concave upward?

The uncertainty about the reptilian parabola actually being concave upward might be illustrated this way. Suppose, among mammals, there survived today only apes, edentates, and insectivores. A curve passing through them would be concave upward, and would be meaningless. When we compute the reptilian parabola on the basis of modern forms, we are dealing with a very fractionated and largely extinct class of vertebrates.

For that matter, this same caution must apply to the average mammalian curve we have computed, although, since the data are so much richer, this curve may carry more value.

In the case of both curves, strictly speaking, they declare simply that, *among surviving forms of the class,* the center of gravity for log brain/log body weight ratio at any given log body size is located at the point $P\ (Y, X)$, and that there is apparently regularity, not whimsy, in the location of these points. The other justification for computing these lines is, that the principles of technique involved would seem to remain valid, even if the raw data were far more complete.

What we need badly, then, is more data, especially reconstructive calculations from paleontology.

Let us repeat once more that, had we more figures for archaic mammals, the modern mammalian constellation would appear as what had been left over from a very extensive erasure of the mammalian block. *Equus* survives, *Eohippus* and *Mesohippus* are extinct. There are two kinds of erasure in the constellation, and their distinctive characters are important in truly evaluating what survives. The ancestor of the titanotheres was transmuted into his descendants, and therefore became extinct only in a transmutational sense. The definitive titanothere became extinct because he ceased to produce progeny. The earliest titanothere ancestor does not appear in FIGURE 18, but in terms of the X, Y axes, quite probably he would be located somewhere to the lower-left, in a portion of the constellation still occupied by extant mammals of entirely other orders. But the titanothere himself would occur in a portion of the field now vacant, as do *Diplobune*, *Uintatherium*, *Anoplotherium*, etc. And when the titanothere existed, undoubtedly the portion of the field which once held his ancestors contained

forms that were to be the ancestors of certain present-day mammals. That part of the mammalian field or block which today is occupied, say, by elk, grizzlies, whales, apes, and man, was probably still virgin and unoccupied. The axis of the mammalian constellation, therefore, may be imagined as having shifted slowly upward, but maintaining something of a fixed point at the lower-left end.

The situation, *mutatis mutandis,* may have been similar in the reptiles. We are not to infer, from the relatively "brainy" place held today by the Crocodilia, that the extreme dinosaurs would have stood even higher. We cannot tell but that the Crocodilia, exceptionally large as they are among surviving reptiles, owe their present salvation to their (for a reptile) exceptionally large brain system.

Farther along, we shall analyze the ratios between the constants b and c, which determine the elevation and curvature—the velocity—of the various parabolic cephalization exponents. Then, the present-day reptilian situation will add a further and rather striking commentary on this total situation.

So, the fidelity of a reptilian parabola curving positively and mammalian parabolas curving negatively has its pros and cons. But there may be neurological justification for the difference. Actually, we are not comparing neurological comparables, since the mammalian curves are very largely taken up with increases in a portion of brain which in the reptiles does not exist. The reptiles, we might say, the "brain stem" animals. Their increase in brain size, from tiny-bodied forms to giant, is probably necessitated almost exclusively by the much more elemental nervous processes than those having their seat in the association tracts of the mammal. The problem of number of motor and sensory fibers per body size can hardly be discussed here, but Ariens Kappers (1929, p. 198) has said, "Brain weight also depends on the size of body. The influence of body size on the nervous system appears already in the greater number of root fibers of spinal nerves in larger animals compared with smaller ones. So the number of the motor root fibers increases with the transverse diameter of the total musculature, while the number of sensory root fibers increases with the surfaces of the body and (proprioceptive fibers) with the transverse diameter of the musculature." When, therefore, we watch a mammalian parabola rising above the reptilian, and notice the falling-off in the log brain weight which, nevertheless, is taking place, perhaps we should discount further another parabola, buried within this mammalian rise, having a positive curvature similar to the reptilian, which would trace the phyletic growth of that part of the mammalian brain most nearly equivalent to the reptilian.

Accordingly, the foundations of the reptilian formula are very fragmentary, the mammalian only less so. We must make due reservations. Nevertheless, all circumstances considered—especially the principles of method involved—it seems permissible to accept what the formulae say. (For a final comment, see heading F 2 a.)

We return now to the behavior of the b and c constants in the sheaf of mammalian parabolas.

F. THE STATISTICAL ANATOMY OF THE MAMMALIAN CLASS (Continued)

THE CORRELATION OF THE CONSTANTS b AND c

1. The Individual Behavior of the Constants among Several Mammalian Stocks*

On previous occasion, we have drawn the analogy of $2c \equiv g$, $b \equiv \nu_0$, in the missile trajectory, where g is gravity, ν_0 is initial or muzzle velocity. But in $Y = A + bX - cX^2$, both b and c have varied per instance. Is the variation regular or random: are b and c correlated or not?

In the case of missile trajectories, the greater the initial velocity (up to an angle of elevation of 45°, or arc tan 1.0000), the greater the horizontal distance covered. The zenith is always half-way over this distance. In the cases of our study, the greater the value of b, the greater, apparently, is that of c, so that the zenith of the highest parabola occurs at the lowest value of X of any parabola.

Before we can be certain of a regular increase (arithmetically) in c as b increases, we must obtain more parabolas. But the less satisfactory condition of the raw data of other orders makes their formulation only a reconstruction. Any conclusions we finally draw must be tempered by this stricture.

We have assumed the right to pass the parabolas of *Homo*, carnivores, and antelope through the field of the insectivores. We have formulated a general mammalian parabola without that assumption, yet it has converged back into the insectivore field. We can hardly be far wrong if we pass other mammalian parabolas through that same region. (To be sure, *below* the insectivore node the parabolas will then have crossed each other and assumed a reverse relationship which hardly represents any true biological situation. Our mathematical devices must stay tempered with common sense.)

In certain cases, we have used also an arbitrary point $Y = -1.900$, $X = 0$. A glance at FIGURES 20 and 22 will tell why. So will the

* The references to "formulae" under this heading are those of TABLE 26

values of A in entries 5, 6, 7, 16, 19, 23 of TABLE 26: we want a value of A close to these, but preferably slightly less, arithmetically, than any or most of them. (In TABLE 26, the least reliable formulae thereby are 13, 15, 18.)

The formulae, then, appearing in TABLE 26 are derived mostly from the same assembly of mammalian data as before, with occasional supplementing from von Bonin. In each case, there have had to be at least 2 genera (*e.g.*, the extinct ungulates from TABLE 13) so spaced that the parabola obtained from them would not be grotesque in comparison with the rest. In certain cases, notably that of the rodents, and perhaps also the Cetacea, the results have seemed grotesque to me anyway (although the rodents are well represented—see TABLE 14), but it would be a biased procedure to omit them. Even if the genera be spaced at good intervals, there is no assurance of fitness in their parabolas. Unless the genera are represented by something like their mean or type values, they will be misleading. If an order is being formulated, the few genera that stand for it may be atypical of the order; that is, in a fuller representation of the order, perhaps they lie at some distance from the true line of mean tendency for the order. Also, it is more difficult to obtain a satisfactory parabola, if the members of the order are all small. If, in addition, their correlation of body and brain weight be poor—a trait that appears to hold among small-bodied orders more than among large-bodied—again the result may be badly "off." This apparent fact may be important, evolutionally; for good results in geometric statistics, it is vitiating.

Once these deterrents are recognized, we may yet find that the aggregate of the evidence points in a certain direction. Hence its use.

The Stocks Severally

Proboscidea. Crile and Quiring give data of 1 male *Heterohyrax*, whence

$$\log \text{brain weight} = 1.0888,$$
$$\log \text{body weight} \doteq 2.8751.$$

If we make an average from 2 Indian elephants (from Brummelkamp, see our TABLE 13), then average that quantity with Crile and Quiring's 1 African elephant, we obtain for the genus:

$$\log \text{brain weight} = 3.7031$$
$$\log \text{body weight} = 6.6628.$$

The use of the hyracoid at the critical mid-range of the parabola is very risky. So, in TABLE 26, formula 12 uses the two points above, plus

the shrews, and formula 13 uses ($Y = -1.900$, $x = 0$) the shrews and the elephants, but not the hyrax.

Perissodactyls. There are plenty of horse data in Crile and Quiring But the horse is a domesticated animal: the body appears to me exaggerated with respect to the brain, when it is compared with the zebra and the rhinoceros. Unfortunately, I have no data on the Przewalski. But the zebra data are good. (Incidentally, the zebra ontogenetic curve comes very close indeed to that of the domesticated horse.) While TABLE 13 has supplied reconstructions of two *Mesohippi*, their use would have given a parabola so very flat as compared with that of the Pecora and Carnivora, and it would have lain so far away from these, that it has not seemed best to place reliance on the reconstructed values of two *Mesohippus* specimens.* The perissodactyls have been formulated twice (formulae 14 and 15), using in both cases the zebras, the rhinoceroses, and the shrews.

Cetacea. In Crile and Quiring's work, and in von Bonin's collection, there are data for the genera *Phocaena*, *Delphinapterus*, *Lagenorrhinchus*, *Tursiops*, *Globiocephalus*, *Megaptera*, and *Balaenoptera*. Von Bonin considers the latter two genera as unusable, because of an excess amount of fat, so he calculates his rectilinear regression without them Just at what point one should stop and say, "From such a body weight on, the animals all have too much inanimate weight to be considered," I cannot say. The problem presents itself whenever we have a sample population of mature animals, such as monkeys, or man, where body weight continues to increase long after cessation of brain growth, or where a digestive tract may be heavily loaded. As another instance, Crile and Quiring's adult horses are mostly old animals (as well as domesticated), hence unusable. Anyway, to return to the Cetacea: formula 21 includes all the whales, taken at face value.

The result is that the values of b and c are both extraordinarily high The isauxon point is far above that of all other mammalian stocks except the primate. This may be the actual situation, but it obviously must be suspect. The curvature of the parabola may have two defects: the lesser whales, which are placed close to the human proportion of brain/body weight, may by that token be immature specimens; the great whales, by reason of much blubber, may be unduly weighted the opposite way The result would be a high-arching, rapidly down-curving parabola. So, a reassessment (formula 22) includes only the great

* There are several similar instances. where a genus or other division places a little too far to the right and not high enough for the rest of a series: *Hyrax*, *Tragulina*, Viverridae, Lemuroidea on the whole: as far as existing evidence goes. Further data in these and other cases are very desirable.

whales, the shrews, and the now familiar point ($Y = -1.900$, $X = 0$). Perhaps the true state of affairs is intermediate to formulations 21 and 22.

Pinnipedia. These include all genera available, plus the point of the shrews.

Edentates, Chiroptera, Amblypods, Anoplotheres. As with the Cetacea, Proboscidea, and the Perissodactyla, the point ($Y = -1.900$, $x = 0$) is assumed. The positions of *Coryphodon* and *Uintatherium* (FIGURE 20) necessitates the use further of the shrew point. In the case of the other three, the parabolas curve close to the shrew point, anyway.

Rodents. These are a special problem. The smaller forms are close to the shrews in body size, yet the brains are larger. In fact, the mice may be even smaller-bodied, yet their brains may be at least as large. Whether the rodent order has had a peculiar and unique morphological evolution behind it or not, at least its proximity to the insectivores on a log-log chart does not tolerate passing a parabola through the shrews, as we have done with orders farther removed in size. Furthermore, we encounter again the phenomenon which seems to allow wide variation in brain weight per body weight among the small mammals. Thus, the Myoidea (FIGURE 14, entries 21–32) trend roughly along a straight-line slope of .56, but scatter widely at the lower-left end of the constellation. This can hardly be altogether an error in weighing tiny brains, for the edentates and marsupials also show low correlation between brain and body weight (see FIGURE 14). This means that we must either omit the rodents from any statistical manipulation that treats the other orders as samples of a series or a population, or else we must include them and consider their discrepancy as a sampling-error, like any other in a given set of data. There is a third possible procedure. Since, in other mammalian orders, we have usually been dealing with only limited parts, a family or a superfamily, etc., we may extract a family or a superfamily of rodents and treat them to a formulation. Whatever we do, we may be sure that the limited character of our data will allow us but a rough picture of what the situation is or has been among these animals. Accordingly, I have evaluated the rodents (formula 8) by the usual method of taking the means of their genera and using these for a mean-squares parabola. The result does not look good. Then the Sciuridae have been formulated separately, this time passing the line through the shrews. The sciurid value for A still is somewhat "off," when compared with the other mammalian lines, but until much more is known about all the stocks, particularly

about their variance, this approximation will have to suffice. At least, FIGURE 23 shows that either formula places the point (b, c) in line with the others.

Marsupials. In no formula summarizing the mammals have we used any but Eutheria. The formulae of the marsupials, therefore, bear comparison with those for the "mammals." Taking all genera, from von Bonin's and our own assemblies together, we obtain a mean-squares parabola (formula 2) that is very flat and more nearly of a shape with that expressing von Bonin's Eutherian data (formula 4). If, on the other hand, we use as one datum the point $(Y = -1.900, X = 0)$ as before, we obtain formula 3. As for how close to the shrews the parabola of either formula passes: when $X = 1.24$, formula 2 gives $Y = -.374$ and formula 3 gives $Y = -.610$; the shrew value being $Y = -.4597$, which is intermediate.

2. The Correlative Behavior of the Constants *b* and *c* in the Mammalian Parabolas

a. The Formula Relating c and b

Consider FIGURE 23 and TABLE 26.

In spite of the various degrees of independence between bodies of data (nothing could be more independent than the reptilian from the mammalian), *all* the points in FIGURE 23, which are the plots of c against b, follow very closely a straight-line tendency. Note that the reptiles are in consistent agreement. (In FIGURE 23A, the scales of b and c are equal, while in FIGURE 23B, the c-scale has been exaggerated 10 times.) So close do the points all come to lying on a straight line, that it hardly matters what procedure we adopt for formulating that line. Of the several possibilities, here are four:

(1) By using the same 10 mammalian parabolas that have yielded formula 6, TABLE 26:

$$c = f(b) = .0697 - .118b.$$

(2) By using the same 10 mammalian parabolas that have yielded formula 7, TABLE 26:

$$c = f(b) = .0884 - .1335b.$$

(3) By using the values of c and b from the reptile parabola (1, TABLE 26) and the mammalian values of c and b in formula 5, TABLE 26:

$$c = f(b) = .0784 - .115b.$$

(4) By using the same reptilian values, and the mammalian of formula 7, TABLE 26:

$$c = f(b) = .0842 - .1305b.$$

The range in slope in these four is from $-.115$ to $-.1336$. Their arc tans. range from approx. $-6° 34'$ to $-7° 36'$, a difference of a little over 1°.

Notice that the slopes in the formula using the reptilian point (the third and fourth) do not differentiate themselves from those using mammalian values only (the first and second).

For the correlation coefficient of *b*, *c* see later.

It is striking to see the reptilian point according with a line determined from the formulae of extant mammals only, and to see that the Anopletheria and Amblypoda (points 17 and 18) likewise seem to accord. The lack of other similar points forbids pushing the speculation farther. If, however, other data should some day make of the present (*b*, *c*) formulation a more widely applicable generalization, then indeed we should be close to some phylogenetic law.

Addendum

A Note on the Reliability of a Positive Curvature to the Reptilian Parabola

On an earlier occasion (see Section E 1), we doubted that a parabola of plus-curvature truly represented the reptilian morphology. We speculated that the Crocodilia might well occupy an exceptionally high position in their class, after the positional analogy (let us say) of the primates among the mammals. Perhaps we are now in a better position to test the crocodilian trajectory.

If we examine the lower-left region of the reptilian constellation (FIGURE 20), we find a cluster of amphibians and reptiles, including the mean point of 15 *Lacertae*. Still farther to the left are several scattered specimens, R 11, A 10, A 12, respectively, Little Gecko, *Hyla arborea*, *Alytes obstetricus*. Very little reflection is needed to conclude that no matter how tiny in body we may suppose any reptile ever to have been, the brain weight could not have been much lower than this region: that is, we cannot imagine a curve coming up to the region under scrutiny from some region far below it on the *Y*-scale. Even fishes (see Brummelkamp, 1939) would not locate far down vertically below the region of R 11, etc. Therefore, we can hardly be far wrong in imagining the ancestral crocodilian trajectory passing through this general region.

Now, if we may invoke the equations relating *c* to *b*, and choose the point of the 15 *Lacertae* (R 4 on the figure) as an at all safe, even if rough, proportion for the ancient crocodilian ancestors, and take the

average values of brain and of body weight for the five *Crocodili* in TABLE 21, we have the following equations:

$$Y = A + bX + cX^2$$
$$-.91721 = A + 1.69897b + 2.8865c \text{ (15 \textit{Lacertae})}$$
$$1.05308 = A + 5.12931b + 26.310c \text{ (5 \textit{Crocodili})}$$

and
$$c = .0842 - .1305b;$$

whence,
$$Y = -1.1538 - .0046X + .0848X^2.*$$

The indication is, therefore, that we cannot pass from the Lacertae proportions to those of the Crocodilia, by way of a parabola having a negative c-value, without doing violence to a reasonable set of proportions between brain and body weight for the crocodilian phylum. In which case, the concept of an increase in *proportion* of log brain-weight as log body weight increases, in at least some reptiles, seems strengthened.

b. Notes on the Degree of Independence between the Constants c and b

When points tend to cluster along a single line (as we have had other occasion to remark), that line tells of a correlation between two variables. Any tendency to diverge to either side of that line (experimental error aside) may be looked upon as bespeaking a degree of non-correlation between the two. Since the points of FIGURE 23 do not lie perfectly on one straight line, the question is in order whether the deviations are those of experimental error only, or whether the several mammalian stocks actually do, or did, tend to vary a little in the ratio between b and c.

To some extent, the error must indeed be one of samplings and of the assumptions made in deriving the formulae. Yet the deviations, I believe, must also be to some extent real, and not just statistical, error. For instance, the anoplotheres and the extant ruminants actually lie in different places in the mammalian constellation. And from FIGURE 23, it appears that they are not far apart in their values of b, but that the anoplotheres had a larger numerical value of c. (In fact, this raises the question of whether, in closely related animals like these extinct and extant ruminants, this may not be the scheme by which they are related and by which they differ. Naturally, it is still impossible to answer the question.)

From the theoretical standpoint, or speaking ideally, there is no reason why what now appears on FIGURE 23 as a sporadic distribution

* Y is a minimum at $X = .0271$. The line is always ascendant over the stretch embraced by empirical data.

of points, should not once have been somehow more regular, and much more populous. That is, it were imaginable that Nature experimented within limits suggested by FIGURE 23, with all or many possible combinations of b and c. While we cannot know whether this really happened or not, the constellation in the figure certainly represents a fragmentation, for very many mammalian lines are now extinct. Of the ruminants, there has been more than one stock-line, but only one (main trend) remains. The situation being what it is, it is useless to ask whether, if we had all stock lines ever produced by the class Mammalia, the mean of them would coincide with the present mean or not. The position of the edentates, amblypods, and anoplotheres suggests that, by recent time, the line of means had shifted somewhat to the right. But whether the shift has maintained the same slope or has changed it, cannot be answered now.

For the present, we must note carefully that the points on the diagram are quite independent of size of brain and body. The primate point, for instance, belongs to the Cercopithecidae quite as well as to *Homo*. This is illustrated by FIGURE 22, where a sample number of the cephalization exponents is drawn.

We now have the four rules:

1. If b is large and c is small, the parabola will be very steep and its curvature gradual;

2. If b is large and c is large, the parabola will start steeply but will curve more rapidly;

3. If b is small and c is small, the parabola will start at a low angle and will curve gradually;

4. If b is small and c is large, the parabola will start at a very low angle and will curve rapidly.

3. The Correlation Coefficient

Empirically, the relation of c to b has been found almost, but not quite, constant.

Now, the points of FIGURE 23 have very unequal value, so that we cannot obtain an unbiased correlation coefficient. Admitting the handicap, and using the points at their face value, we may compute a correlation. Almost the *poorest* correlation possible would use from TABLE 26 the formulae 1, 2, 4, 8, 10, 11, 12, 14, 16, 17–21, 23, which render:

$$r_{bc} = -.912.$$

A theoretical system of cephalization coefficients, in which

$$r = -1.000,$$

is always possible. Whether it would have empirical backing, or whether the latitude allowed by an imperfect correlation represents something biological, is, of course, impossible to say.*

G. THE STATISTICAL ANATOMY OF THE MAMMALIAN CLASS (Concluded)

ON THE POSITION OF THE CONSTANT A IN THE CEPHALIZATION EXPONENT

1. The constant A in the cephalization exponents has been the value of Y where $X = 0$. This is true, also, in the formula $Y = A + bX$. The Dubois school has called antilog A, which is a in $y = ax^b$ (to use our own terms), the cephalization coefficient.

The scale of a or A is as arbitrary as Fahrenheit or Centigrade, since there is no significant biological point at $X = 0$ (or $x = 1$).

In a previous heading, we have found the intersection between the reptilian and a mammalian parabola It would be more substantial, biologically, to take such a point as the origin of all mammalian parabolas, even though this mathematical quantity is sharper than the biological event it seeks to locate.

With some such fixed point, it becomes possible to symbolize, in imagination, the rise of the mammalian curves out of the reptilian in this fashion:

Let the reptilian parabola be a flexible steel rod, secured at the calculated point of intersection with the mammalian parabola. Then apply a lifting-force to it, close to the fixed point, and also a weight. Gradually increase the upward pull, and at the same time slide the weight farther to the right.

---

* Assume that $r = -1.000$ and calculate a cephalization exponent for *Simia* by assuming only the empirical means Y, X for *Simia* and for *Blarina*, and any one of the formulae $c = f(b)$ Repeat, replacing *Simia* with *Homo*. Imagine, then, a time when the (common?) ancestors of man and of orang had the general proportions of Cercopithecidae It does not matter whether one pretend that the ancestors had already differentiated towards their descendants or whether they were still an unidentifiable portion of the cercopithecid magma. In either case, there could have been statistically a variation in brain weight percentage The variation is present even today and it is wide. Next, take some convenient cercopithecid body weight—say, the mean. Calculate Y for this mean for the cephalisation exponents of *Homo* and of *Simia*. None of the empirical formulae $c = f(b)$ listed in Section F 2a can be made to give values of c and b such that calculated Y's for mean cercopithecid X do justice to the respective exponents and yet fall within the range of cercopithecid brain weights for that body weight.

1 Assume $c = .0842 - .1805\ b$; then,
Simia: $Y = -2.2443 + 1.5918\ X - .1234\ X^2$
Homo: $Y = -3.0338 + 2.359\ X - .2238\ X^2$
Let $X = 3.3021$; then, from
Simia: $Y = 1.6677$; antilog $= 46.5$ gm
Homo: $Y = 2.3162$; antilog $= 207.1$ gm.
2. Assume: $c = .0697 - .118\ b$; then
Simia: $Y = -2.1059 + 1.4625\ X - .1028\ X^2$
Homo: $Y = -2.7331 + 2.042\ X - .1711\ X^2$
Let $X = 2.3021$; then from
Simia: $Y = 1.1671$; antilog $= 14.7$ gm
Homo: $Y = 2.1499$; antilog $= 141+$ gm.

Such conflicting results do not, of course, disprove the possibility of $r_{bc} = -1.000$ However, pending future corrections which may so indicate, it is best to assume the correlation is not perfect. Above all, it is essential to remember the "as if" character of the cephalization exponents.

At every point except where it is fixed, the steel rod will rise, its inclination will steepen, and the curvature will become sharper. Its behavior resembles that of a fishing-rod after a "strike."

In analytical geometry, this behavior is a "transformation."

As a working case, let us transform the reptilian parabola into several of the mammalian of TABLE 26: entries 9, 16, 19, 20, 23, and the general mammalian to which they contributed, 7

1. The fixed point, the point of intersection of reptilian and mammalian lines, is for entries 1 and 7, TABLE 26:

$$Y = -1.6217$$
$$X = .2385.$$

The point is now the origin of a new set of axes:

$$Y' = 0,\ X' = 0. \quad \text{Then,}$$
$$Y' = Y + 1.6217$$
$$X' = X - .2385.$$

The six mammalian parabolas are to be corrected, or "warped," to pass through this point.

This may be done as follows: Take arbitrary values of X, and calculate Y for each; rewrite X and Y in terms of X' and Y'; form the new equations $Y' = 0 + b'X' + c'X'^2$; and solve for b' and c'. The results are in TABLE 27.

A comparison of TABLES 26 and 27 will show that the "corrections" have not been severe For instance, if we calculate primate values of Y', X' from TABLE 27 and derive from them values of Y, X, we obtain TABLE 28.

Two profitable operations now follow in the next two sections.

2. The vertical distances between the reptilian parabola and any mammalian are given by:

$$Y' \text{ (mammal)} - Y' \text{ (reptile)} = (b'X' - c'X'^2)_M - (b'X' + c'X'^2)_R.$$

Rewriting the two sides of the equation,

$$RY = mX' - nX'^2.$$

Interpreted back into the antilogarithms, it reads

$$\frac{y' \text{ (mammal)}}{y' \text{ (reptile)}} = \frac{\text{antilog } (b'X' - c'X'^2)_M}{\text{antilog } (b'X' + c'X'^2)_R}.$$

Writing the ratio of the y's as ry,

$$ry = \frac{x'^m}{x'^{2n}}.$$

See TABLE 29 for the values of RY.

The regularity of the family of curves suggests that m and n are correlated. (With the Dubois formulation, this could not happen,

since all slopes are parallel.) The ratio of n to m is a very good straight line for Sciuridae, Pecora, Carnivora; Pinnipedia: $n = -.05413 - .0773\ m$; *Homo* deviates. In this scheme (it has not seemed worth illustrating by a graph), the "lowly-brained" Sciuridae are placed at the left end of the line, the "higher-brained" to the right, with *Homo* leading, followed by the Pinnipedia and then by the Carnivora and Pecora simultaneously. The line $RY = mX' - nX'^2$ is a parabola giving 0 values where the mammalian cephalization exponents intersect the reptilian, and giving a maximum value where

$$X' \ (\textit{mutatis mutandis},\ X) = \frac{b_M - b_R}{2c_M + 2c_R}.$$

TABLE 30 gives these maxima twice: once as calculated by using TABLE 26, once by taking the first derivatives of the formulae in TABLE 27. In the latter case, the point X' so found has to be reinterpreted into $X = X' + .2385$. It will be seen that agreements are very close, again indicating that no great violence has been done by bending all curves to pass through a common point on the reptilian line.

These values of X' happen to occur in the most populous portion of the mammalian constellation (see FIGURE 20), where (except for Primates) body weight ranges around 6–10 kg. The value of X' in the primate line happens to occur where man is located. Whether these are accidents and devoid of significance, I cannot say.

3. The transformation now resolves itself. The X' axis remains unaltered, but the mammalian lines have come about by a change of *slope* λ from the reptilian. This is a one-dimensional strain: a transformation in the Y'-dimension only. Then

$$\lambda\ (\text{mammal}) = l \times \lambda\ (\text{reptile})$$

$$\frac{\lambda\ (\text{mammal})}{\lambda\ (\text{reptile})} = l.$$

Since
$$\lambda = \frac{dY'}{dX'} = b' + 2c'X', \text{ we have}$$

$$l = \frac{b' + 2c'X'\ (\text{mammal})}{b' + 2c'X'\ (\text{reptile})}.$$

The curves of empirical l are formulated in TABLE 31.* Notice that l is always great for small values of X', and falls off (parabolically) as X' increases. That is, the distortion or transformation takes place most strongly where the mammals are smallest of body.

A clue to quantitative understanding of evolutional changes in body

* The high arithmetic values of the constants in the pinniped formula may indicate a faulty reconstruction of its original cephalization exponent.

shape, as well as in the matter of brain/body ratio, lies in the geometry of transformations. D'Arcy Thompson (1942, pp. 1051 *et seq.* The material is present also in the first edition) has depicted it graphically for the skull and the body shapes of certain animals. He has presented a challenge to the students of human evolution which, some day, must be met in terms of a kinetic somatometry.

FINALE

There appears to be nothing aberrant about the rise of man, no need to search for some *tertium quid* by which a whimsical Nature struck off at a new tangent to produce a cerebration which caricatures that of the mammalian rank and file. That *tertium quid* does not exist. In fact, if one is looking for a really great upsurging of cephalization, one should contemplate the long and steep rise (of the logarithms) from Reptilia to Insectivora, or the lesser, yet still enormous, achievement from Insectivora to Hapalidae. Compared with these, the rise from monkey to man is the smaller biological achievement. Let us not confuse standards. The *practical results* that come from the exercise of the aggregate human brain, the combined efforts of humans which we term "culture," certainly are spectacular when compared with the analogous achievements of other animals. However, in the world of biological operations, especially when scaled in its intima by logarithms, the abysmal discrepancy does not exist.

As for the word "aberrant," if one feels he must use the term at all, when speaking of mammals, it were well first to be sure just what is meant by it.

SECTION IV

ONTOGENY AND COMPARATIVE ANATOMY

A. PRIMATES

Consider FIGURE 24.

The two fetal human and the fetal *Semnopithecus* curves of TABLE 12, plus their mean (see SECTION III, 3); the postinfantile curves for man, orang, chimpanzee, *Semnopithecus,* and *Rhesus,* from TABLE 11; the exponent of cephalization from the shrew-monkey-man line; and the adult means of species from TABLE 16, are all plotted.

It will be recalled that the comparative-anatomic line was constructed on the assumption of mean values for shrews, for Cercopithecidae minus the baboons, and a heterogeneous congeries of 41 humans.

1. The chart shows a marked parallelism between the fetal lines, derived from an absolutely different mass of data, and the comparative-anatomic, over a considerable stretch. The lower ends of the two lines diverge, and perhaps we shall recall the suggestion that it is apparently before our data come in that the embryo has pushed up its brain weight to a size greater than that of a putative adult ancestor of corresponding body weight. The embryo has appropriated a larger percentage of its growth-energy to the brain.

2. Most, but not all, of the adult monkeys cluster about the comparative-anatomic line. This, of course, is by hypothesis. The exceptions are an interesting assortment: the great apes, the gibbons, the *Rhesus* macaque (according to Zuckerman and Fischer, but contrary to Spitzka), most of the Prosimiae. The variability of body weight is responsible, but far less so that of brain weight. Some of this variation may have positive significance, some of it may not. Thus, there can be no doubt about the placement of the great apes; the several species of gibbon corroborate each other; so do the lemurs. The Cercopithecidae, on the other hand, are uncertain because of the macaque contradiction. We have the ontogenetic data for *Semnopithecus* and *Rhesus*, which certainly do not permit pushing these forms back among the clustered means of other Cercopithecidae. We could probably fill out the interval between them all if we had more data. (Some, but by no means all of the adults are marked as emaciated. Spitzka was very careful

to annotate his juveniles, and so were Hrdlicka and Crile and Quiring. Thus, the safest procedure was, as noted before, to use the means of the aggregate adults.)

We remark again that the baboons, as well as the Hapalidae of Hrdlicka and of Crile and Quiring, immediately align themselves on this same curve. Interestingly enough, from the standpoint of the assumption that the smaller members of a stock should be the more conservative, so does that minute prosimian, *Microcebus minimus* (to judge from the single specimen listed by von Bonin). So, taken altogether, it seems a not unfair conclusion that the *Homo* line is not unrepresentative of the rise of the Primates, and that there is a phyletic tendency to push out to larger body size, once a certain brain level has been reached, without much further increase in log brain weight.

At any rate, it is very doubtful if any fetal lines of monkeys would lie above and to the left of the human fetal line, since that would give them, shortly before birth, a brain size larger than that of a human fetus of equal body weight, but of far less advanced morphology. This should make the human fetal line a limit of an area; the postnatal growth of the monkeys must take place between it and the point where their adult forms are registered. If they all should have postinfantile slopes of the general range between *Semnopithecus* and male *Rhesus* (within this range have occurred all animals we have dealt with, except man and the anthropoids), then we may form rough ideas of where the several monkey forms would have their fetal development.

3. The human ontogenetic line turns out approximately parallel to the primate line. Measured in terms of logarithmic cycles, the horizontal distance between the two lines is approximately uniform over a large part of the inscribed area. Man actually travels a very short horizontal distance from the fetal slope to his position on the line of pseudo-phylogeny. So do some monkeys. Apes travel a very long distance. So do large baboons, like *Hamadryas*. These forms I would term "giants," at their level of brain size.

B. PRIMATES AND OTHER MAMMALS

1. A. H. Hersh (1941, p. 138) publishes a diagram furnished him by B. G. Anderson, which is a "hypothetical logarithmic plot intended to illustrate ontogenetic and evolutionary relative growth and the way in which they may be related to each other." I had already been working on FIGURE 29 when this diagram came my way, and I was struck by the coincidences. Hersh's diagram has a set of steep ontogenetic lines which take a sharp turn to the right by reducing their slopes, and an-

other, a "phylogenetic" set, which start with low slope and turn abruptly upward at much increased slopes. The two sets of lines, therefore, inevitably intersect. In case of the lines developed as a system in the present essay, a phylogenetic claim is not being made for them, as we have been careful to state. None the less, the similarity of scheme or structural system cannot but be noticed.

About the lower ends of my own curves, nothing but surmises could be offered. But, at least after the data begin, the upward sweep of both sets of lines is clear enough. My empirical lines differ from Hersh's hypotheticals in certain regards: his are mainly rectilinear; my comparative-anatomic lines are parabolas (because of the c-factor); and my fetal lines are plotted as straight lines only, as the best approximation it is safe as yet to make. In the matter of brain and body weight, I am not willing to stress the "law of allometry" as strongly as is done by Hersh. Nevertheless, we seem to agree on something more fundamental than a mere formula.

2. In FIGURE 24, I have confined the representation to three orders: Primates, Carnivora, Artiodactyla. On the fetal side, the Carnivora are represented only by the (domestic) cat.

I trust I court no misunderstandings in assembling these materials.

The artiodactyls are represented, on the fetal side, by the Bovidae of Section II; on the postnatal side, by the antelopes of Section III.

It is noticeable that the domestic forms—cat, sheep, ox—reach adult size below the comparative-anatomic lines.

The crosses are plots from the line of mean mammalian, comparative-anatomic trend.

How far the ontogenetic lines parallel the comparative-anatomic, can be seen at a glance. At their lower ends, the ontogenetic trail off markedly from any parallelism, but we cannot say whether they should continue (in the backward direction) to trail off, or whether they should dip again.

Another striking feature is the very small interval between an ontogenetic line for Carnivora or Pecora and the comparative-anatomic, as contrasted with the very wide interval of the Primates. This ties in with the much more gradual slopes, both ontogenetic and comparative-anatomic, of the less cephalized orders. Without that coincidence, there is no geometric reason why the cat or ox fetal curve should not lie as far from the comparative-anatomic curve, during its earlier career anyway, as in the case of the Primates. But what would be the consequence? A fetal specimen with a far huger brain than it actually possesses, but an adult brain of the size actually obtaining.*

* Analogous to speeding very quickly towards a destination, then slowing down to a crawl in order not to arrive too soon.

3. The small squares on the diagram indicate the spot where, in the pertinent formulae of TABLE 26, the slope is 45°: *i.e.,* the point of isauxon. Before that spot is reached, the log brain weight is increasing faster than the log body weight. It was to have been expected that, for the mammalian cephalization to ascend out of the reptilian, the log brain weight would gain over the log body weight. The situation, as we have seen, boils down to this:

(a) The log brain weight usually ceases to gain over the log body weight, somewhere in the neighborhood of the Insectivora. By token of the analysis in Section III F, the exponent with the lowest value of b and of c (numerically) will reach isauxon the earliest. The Primates owe their preeminence to a particularly advantageous ratio of b to c; although, in principle, there is nothing peculiar about this ratio.

(b) Nevertheless, cerebral complexity continues to increase, as we know. But it does seem reasonable that the "set" of these parabolic directions was a necessary prelude or prerequisite to the evolution of high cerebral "capacity."

(c) The primate fetal curve stands off at a far greater elevation from the comparative-anatomic than in the other orders. Both curves are steeper than those of other orders. This double steepness must be ontogenetic and comparative-anatomic aspects of fundamentally one and the same thing.

(d) It may be that primate species are far more prone to vary in cephalization than the other orders. This point would, at least, be consistent with the preceding ones.

4. What has now happened to the *"exposant de relation"* and the *"coefficient de céphalisation"* of Dubois *et al.?* The criticisms of von Bonin (that the premises on which they are based are now archaic), seem quite justified. Yet von Bonin, I believe, went too far when he jettisoned the promise of dynamic implications below the face of his chart. The concept of a "progressive cephalization," I am convinced, is real. This shines through Dubois' exponent and coefficient. It persists in the issues underlying Brummelkamp's illustrations which we have reproduced in this study. This is despite of what I feel to be a fact: namely, that the sloping staff of straight parallel lines in those illustrations really extract just about what they first put in. Repeatedly, we have seen the slopes of the parabolas, in their mid-regions, ranging their values in the neighborhood of .56. In those limited reaches, then, the cephalization coefficient still remains as a handy rule-of-thumb for comparing the relative heights of several lines of cephaliza-

tion. That the cephalization exponent herein set forth comes, as I really believe, closer to describing the situation than does the *exposant de relation* and the *coefficient de céphalisation,* in no wise alters my conviction that Dubois' first "cracking" of a very important problem is a landmark in the progress of biology.

ADDENDUM

Since finishing this paper, an old reference in my card catalogue has turned up. It is a quotation from M. Boule, *"La Paléontologie Humaine en Angleterre,"* p. 67: *"Un jour viendra où l'on découvrira un Hominien de petite taille, à la station à peu près droite, à la boîte cérébrale relativement très volumineuse par rapport au volume total du corps, mais très inférieure, en valeur absolue, à celle de tous les Hominiens déjà connus. Ce sera le véritable* Eoanthropus."

SECTION V

A. SUMMARY

1. In man and other mammals, relative increase in brain and body weights, in ontogeny, has three periods after the embryo has become a fetus: (a) a fetal, measurable as a *first approximation* by $y = ax^b$, or $\log y = \log a + b \log x$, where y is brain weight, x is body weight, in grams; (b) a postinfantile, measurable in like terms, but with a far higher $\log a$ value and far lower b value; (c) a transitional period between them, starting at an indefinable point before birth and ending indefinably in the postinfantile period. This third period is an adjustment between the two others, and in its logarithmic form is curvilinear. The first derivatives of the rectilinear formulae under (a) and (b) form upper and lower asymptotes, to which the first derivative of (c) is an asymmetric logistic:

$$\frac{d(\log y)}{d(\log x)} = \frac{K}{1 + e^{v}} + d,$$

where d is the lower asymptote, K is the difference between upper and lower asymptotes, and $v = f\ (\log x)$, which is expressed fairly as a third-degree parabola. However, as far as can be ascertained,

$$\int\left[\left(\frac{K}{1 + e^{v}}\right) d(\log x)\right]$$

is unobtainable. Hence, the integral curve remains a matter for mechanical methods.

Quantitatively, the sexes in man never behave alike throughout growth.

The comparative behavior of man, other primates, and certain ungulates, carnivores, *et al.* has been sufficiently summarized in Section IV B. It will be seen there that the differences of behavior lie in the values of the constants. However, "first approximation" (above) has been used advisedly, since the fetal curves perhaps are straight lines only in a general way. More accurately, they may fit some as yet unformulated parabolae of not less than third degree, or some other undulant curves. If so, it is logical to standardize them for comparisons, by way of a common scale of logarithmic cycles.

2. On the side of comparative anatomy, evidence is presented for

an *exponent of cephalization,* by which route a line of mammals may be considered progressively to have raised its brain weight y in proportion to its total body weight x: where $Y = \log y$, $A = \log a$, $X = \log x$,

$$Y = A + bX - cX^2,$$

$$y = ax^{b - c\log x}.$$

The last formula, it is considered, is more correct in the present case than any form $y = ax^b$, which by various authors has been applied as the "law of allometric growth" to many different biological phenomena. It means that we should write, not:

$$\frac{dY}{dX} = b, \qquad \text{but:}$$

$$\frac{dY}{dX} = b - 2cX,$$

$$\frac{d^2Y}{dX^2} = -2c,$$

in which c is an important term. By analogy, $2c$ is compared to the gravitation constant g; except that c is a variable, which is therefore analyzed further (see 3, below).

The formula is at variance with that of Dubois, which, in the symbolism above, would read $y = ax^{.56}$, the latter being the basis for saying that, when two different but related species or genera of whatever body weight have equal values of a, they may be joined by a line $Y = A + bX$, where $b = .56$.

While agreeing with some of von Bonin's criticism of Dubois, the parabolic formula above does not agree with von Bonin's conclusion that regression of log brain weight on log body weight for the Mammalia taken as a whole is rectilinear. The reasons for the disagreement are given.

As a secondary matter, the conclusion of Lapicque, that different-sized strains or individuals within a species may be arranged along a logarithmic diagonal of slope ranging to either side of .25, so that this quantity may be taken as standard, is not justified by the present study. The variations on either side of that figure are characteristic of the data used in each case, and cannot be dismissed. That adults within a species tend to orient along an axis of a positive slope that is steeper than that of the mass tendency in growth of the species, is just another phenomenon of variance in a species. It does not justify such artificialities built upon it as the Index of Cerebral Value of Anthony and Coupin.

3. The anatomy of the relationship between b and c is as follows:

$c = f(b)$, and is rectilinear (see FIGURE 23);

$r_{bc} = -.912.$

Therefore, a system exists by which may be obtained a set of second-degree parabolas, each parabola being the cephalization exponent for a given evolutionary line of mammals, as stated above.

The set, as a whole, has a parabola of mean tendency, and an expressible variance for each constant in the formula. Thus, each parabola of the set can be measured absolutely and relatively for its deviation from a central norm.

The text explains:

(a) TABLE 26, which lists 23 second-degree parabolic cephalization exponents, covering the Reptilia as a class, the Mammalia as a class, and various mammalian evolutionary stock-lines (see FIGURE 22).

(b) The variance along the straight line, $c = f(b)$, is constant, in that the errors on either side of that mean are about the same over its entire ascertained length. This variance may be measured on a line intersecting $c = f(b)$ at any point and normal to it.

4. By a very slight correction in the curvature of the parabolas which does no significant violence to the data, they can all be warped to pass through a common point of intersection with the reptilian parabola. Thereby, it becomes possible to shift the axes so that this point shall be the origin. This makes it possible to measure at all points the vertical difference between any mammalian line and the reptilian; and, further, to calculate the transformation of slope from the reptilian to that of any mammalian order. This is done by considering the elevation from the reptilian parabola to the mammalian as a distortion of ordinate. The ratio between the slope of a mammalian parabola and the reptilian is a second-degree parabola.

In retrospect, the author is aware that many criticisms can be leveled at the steps of this development. The data are not too abundant; their sources are varied and not subject to uniform control; the shrews are made to bear the weight, it seems, of the entire mammalian class; sometimes much is made out of little, etc. Of course, it is too late now to apologize. This study has not been written for the individual who contends that one has a right to open his mouth only after the mass of data has reached overwhelming proportions.

As for the numerical validity of the several formulae, quite probably every single one of them will be subject to correction when more material becomes available. As they stand, I think they are good enough to demonstrate a certain regularity of behavior of all mammals, including man.

Another weakness remains. We have postulated an exponent that is

a parabola of the second degree. We have not disproved the fitness of some other curve that is concave downward. We have gone ahead to the dissection of the b and c constants, as though they had been definitely established as entities. I think it is true, nevertheless, that the configurations shown in FIGURE 22 support them as a working hypothesis, for the curvatures of the various lines are the most abrupt for the highest lines, pointing to zeniths reached the soonest. Certainly, also, the tie-in of the reptiles that has been suggested at least makes for a system. The adoption of the second-degree parabola has had a certain logic of the immediate. Hitherto, authors have exploited the formula (quite as empirically determined as our own, and with far less analysis to back it), $y = ax^{.56}$ or $Y = A + .56X$. It is quite definitely true that such lines demand a negative correction-term, if they are to fit a more lengthy range of data. It lies immediately at hand, therefore, to try $-cX^2$. The results have been good, but, admittedly, our approach has the weakness of its strength: it is purely empirical. I fail to see how it could be otherwise; nevertheless, there is much to be done before the hypothesis becomes secure.

However, it can hardly be denied that extant animals can be related to each other by some such curves as we have demonstrated It is also undeniable that the larger-bodied animals budget a smaller percentage of their total weight-producing growth energy to brain than do the smaller animals. The curvatures that run between or through them must be concave downward. In brief, I have more confidence in the scheme of principles set forth herein than in any of the concrete numbers that have served to bring them out.

B. CONCLUSIONS

Ontogeny

1. In man and in other mammals, the growth of brain weight with respect to body weight has three periods: a fetal period, a transitional period through infancy, and a period thence to adulthood.

(In the following, all quantities are referred to in terms of their logarithms.)

2. In most of the fetal period, the growth plots approximately a straight line. In life after infancy, it plots another straight line. The first is steep and the second very gradual. Between the two is a period shown by a curving line that joins them.

3. Whatever be the meaning of this, it is during the time of curvi-

linear growth that mitoses of brain cells diminish to none, so that growth of brain material, thereafter, occurs only by enlargement of cells.

4. The sexes do not coincide exactly at any point.

5. When man is compared with some other primates, and the comparison is then broadened to include some ungulates and certain others, it appears that:

(a) In fetal life, the primates have, on the average, steeper slopes of growth than the artiodactyls. But steepness of slope in the fetal period cannot be equated simply with greater adult brain, since the pigs have the steepest slope of the artiodactyls used, and it may be as steep as the human and steeper than the monkey's. But man's line also is steeper than the monkey's, probably on a par with that of the chimpanzee. Superiority in brain size is partly traceable back to the prefetal period.

(b) In postinfantile life, the rise of brain weight, relative to body weight, is less in man than in the monkey. Other mammals resemble the monkey in this respect, except for the anthropoid apes, which distinctly belong with man. In fact, they surpass man in having a *still lower* rise of brain weight relative to body weight.

The artiodactyls differ as much intramurally in their prenatal slopes as the primates. After infancy, the primates are variable as just described, while the artiodactyls (with but small difference in brain organization) are very uniform.

6. The human brain, during the measurable fetal period, does not grow much more rapidly than the monkey's. There are two reasons for the preponderance of the human brain:

(a) *Before* the measurable period, it has already achieved a larger proportionate size. After that, it maintains its lead by but little more than pacing the monkey rate.

(b) More important, the preponderance is in terms of a steeper increase in body and brain taken together with respect to morphological levels of maturation Thus, when the monkey is ready to be born, the human body of the same log weight and only slightly greater log brain weight is still unready for birth. Growth continues for some time thereafter, at a rate which approximately had been followed by both man and monkey. Brain preponderance demands large absolute body size.

7. In man and in other mammals, the adults of different sizes do not arrange themselves at all as though they were merely cessations of growth at various points along the line of mean tendency for the group.

Thus, small adults do not appear like "arrested adolescents," nor large ones as "overgrown" along the same line. This interpretation is at variance with that of Lapicque, who first pointed out the phenomenon.

Comparative Anatomy

8. Phyletic increase in complexity of cerebral organization (its character is not necessary to the discussion) requires phyletic increase in absolute brain size. That increase is contingent upon phyletic increase in total body size, quite as certainly as in the matter of ontogenetic increase.

9. The phyletic increase of brain size, concomitant with that of body size, does not follow the system developed by Eugène Dubois and his followers. On a log-log chart, that system would be a mass of straight lines having an average slope of .56 Instead, the logarithm of brain weight traces against the logarithm of body weight an ascending curve, concave on the lower side

Along this curve occur the various-sized and variously evolved extant families and genera of the order. Such a curve is expressible as a second-degree parabola.

This parabola is the "exponent of cephalization" The principle applies to individual orders or to the general trend among extant mammals as a class. The parabolas behave so as to rise and disperse from a common center in the region of the insectivores.

10 The exponents of cephalization are not random in their constants: there is a clear system of relationship between them. This system has a number of formulable features.

11. The reptilian cephalization exponent is an ascending parabola, concave on its upper side. The constants in its formula fit, nevertheless, into the relational scheme that answers for the mammals.

12. The mammals have arisen from very small reptiles by a formulable step-up in the proportion of body weight devoted to brain.

13. However, the mammalian step-up is largely confined to those portions of brain practically unrepresented in the reptiles. There seems to be no great discrepancy between the reptilian brain/body weight ratio on the one hand, and on the other, the ratio between body weight and the roughly homologous portion of brain in the mammal, when the mammal is equivalent to the reptile in body size.

14. In the total mammalian system of parabolas, the steepest is that primate line which leads to man. But there is nothing at all unique *about it; man is not aberrant. His blueprint, so to speak, was in-*

herent at the time when some primitive animals became mammals, just as the blueprints of all other mammals were inherent. In fact, from cercopithecid to *Homo* is less of a stadium than that from hapalid to cebid or from insectivore to hapalid.

15. When, finally, ontogenetic growth and comparative-anatomy lines are compared, as far back as the former can be traced in the materials at hand, a certain parallelism exists between fetus and comparative anatomy. The fetal line, however, tops the comparative-anatomic, in such a way that a fetus of a given body size always has a heavier brain than some extant adult relative of equal body size, who presumably is less evolved. In the primates, this preponderance is much greater than in any other line. Since, in all this discussion, the ratios are those of a part of the whole, it means that, when one weighs an assortment of mammalian fetuses, for any particular and equal body weight, the primate has put more material into the weight of brain than has any other mammal. And the shapes of the various curves indicate that fetal preponderance starts much farther back in the embryonic life than the measurements extend. After birth, the ontogenetic line has no comparative-anatomic implications now discernible.

REFERENCES CONSULTED

Allen, Ezra

1912. The cessation of mitosis in the central nervous system of the albino rat. J. Comp. Neurol. **22**: 547–568.

Anthony, R.

1928. Anatomie comparée du cerveau. Doin. Paris.

Anthony, R., & F. Coupin

1925. L'indice de valeur cérébrale. Revue Anthrop. **35**: 145–51.

1925–1926. Introduction à l'étude du développement pondéral de l'encéphale. L'indice de valeur cérébrale au cours de l'évolution individuelle: 483–500. Spomenica o počast Prof. Dru. Gorjanović-Krambergeru. Zagreb, Yugoslavia.

von Bonin, G.

1937. Brain-weight and body-weight in mammals. J. Gen. Psych. **16**: 379–389.

Boyd, E.

1942. Outline of Physical Growth and Development. Burgess Publishing Co. Minneapolis.

Brummelkamp, R.

1939a. Das sprungweise Wachstum der Kernmasse. Acta Neerlandica Morph. Norm. et Path. **2**: 177–187.

1939b. Das Wachstum der Gehirnmasse mit kleinen Cephalisierungssprüngen (sog. $\sqrt{2}$-Sprüngen) bei den Rodentiern. Acta Neerlandica Morph. Norm. et Path. **2**: 188–194.

1939c. Das Wachstum der Gehirnmasse mit kleinen Cephalisierungssprüngen (sog. $\sqrt{2}$-Sprüngen) bei den Ungulaten. Acta Neerlandica Morph. Norm. et Path. **2**: 260–267.

1939d. Das Wachstum der Gehirnmasse mit kleinen Cephalisierungssprüngen (sog. $\sqrt{2}$-Sprüngen) bei Amphibien und Fischen. Acta Neerlandica Morph. Norm. et Path. **2**: 268–271.

1939e. Schädelkapazität und Körpergrösse bei den verschiedenen menschlichen Rassen und Bevölkerungsgruppen. Acta Neerlandica Morph. Norm. et Path. **2**: 360–378.

1939f. Über den Zusammenhang zwischen Schädelkapazität und bestimmten Femurmassen, zugleich ein Beitrag zur Cephalisationsfrage von *Pithecanthropus erectus*. Acta Neerlandica Morph. Norm. et Path. **2**: 379–400.

Crile, G. W., & D. P. Quiring

1940. A record of the body weight and certain organ and gland weights of 3,690 animals. Ohio J. Sci. **40**(5): 219–259.

Cruikshank, J. N., & M. J. Miller

1924. The weight of fetal organs, etc. Child life investigations of the Medical Research Council. Great Britain Privy Council. Special Report Series **86**: 63–64.

Donaldson, H. H.

1895. The Growth of the Brain. Scribner's. New York.

1908. A comparison of the albino rat with man in respect to the growth of the brain and of the spinal cord. J. Comp. Neur. & Psych. **18** (4): 345–392.

1910. On the percentage of water in the brain and in the spinal cord of the albino rat. J. Comp. Neur. & Psych. **20**: 119–144.

1911. Studies on the growth of the mammalian nervous system. J. Nervous & Mental Diseases **38**: 257–266.

1916–1917. Growth Changes in the Mammalian Nervous System. The Harvey Lectures **12**: 133–150. Lippincott, New York.

1918. A comparison of growth changes in the nervous system of the rat with corresponding changes in the nervous system of man. Proc. Nat. Acad. Sci. U. S. **4**: 280–283.

1924. The Rat. Wistar Inst. Anat. & Biol. Mem. **6**. Philadelphia.

1925. On the effect of captivity or domestication on the brain weight of some mammals. Wistar Inst. Anat. & Biol., Am. J. Physical Anthrop. **8**: 352–353.

1925. The significance of brain weight. Arch. Neurol. & Psychiat. **13**: 385–386.

1932. The brain problem in relation to weight and form. Am. J. Psychiat. **12**: 197–214.

Donaldson, H. H., & S. Hatai

1911. A comparison of the Norway rat and the albino rat in respect to body length, brain weight, spinal cord weight, and the percentage of water in both the brain and the spinal cord. J. Comp. Neur. & Psych. **21**: 417–458.

Dubois, E.

1897. De verhouding van het gewicht der hersenen tot de grootte van het lichaam bij zoogdieren. Verh. Kon. Akad. Wetenschappen Amsterdam **5**: 10.

1898. Über die Abhängigkeit des Hirngewichts von der Körpergrösse beim Menschen. Archiv. f. Anthrop. **25**(4): 123.

1913. On the relation between the quantity of brain and the size of the body in vertebrates. Verh. Kon. Akad. Wetenschappen Amsterdam **16**: 647 ff.

1914. Die gesetzmässige Beziehung von Gehirnmasse zur Körpergrösse bei den Wirbeltieren. Zeitschr. Morph. Anthrop. **18**: 323–350.

1918. On the relation between the quantities of the brain, the neurone and its parts, and the size of the body. Verh. Kon. Akad. Wetenschappen Amsterdam. **20**.

1923a. Phylogenetic and ontogenetic increase of the volume of brain in the vertebrata. Proc. Kon. Akad. Wetenschappen Amsterdam **25**(Sect. sci.): 230–255.

1923b. Phylogenetische en ontogenetische toeneming van het volumen der hersenen bij de gewerbelde dieren. Verh. Kon. Akad. Wetenschappen Amsterdam **31**(6): 307–332.

1924. On the brain quantity of specialized genera of mammals. Verh. Kon. Akad. Wetenschappen Amsterdam **27**.

1934. Phylogenetic Cerebral Growth. Congrès International des Sciences d'Anthropologie et Ethnologie: 71–75. Royal Anthropological Institute. London.

Dunn, H. L.

1921. The growth of the central nervous system in the human fetus. J. Comp. Neurol. & Psych. **33**: 405–491.

Frechkop, S.

1927–1928. Remarques sur le poids du cerveau chez les mammifères. Ann. Soc. Roy. Zool. Belgique **58**: 109–116.

Harris, H. A.

1929. A preliminary note on the relation of skeletal ossification in the hind

limb to the index of cerebral value of Anthony and Coupin. J. Anat. (London) **63**: 267–276.

Hauger, O.

1921. Der Gehirnreichtum der Australier und anderer Hominiden beurteilt nach ihrem Skelett. Anatom. Hefte **59**(179).

Hersh, A. H.

1934. Evolutionary relative growth in the Titanotheres. Am. Naturalist **68**: 537–561.

1941. Allometric growth: the ontogenetic and phylogenetic significance of differential rates of growth. Third Growth Symposium: 113–145. Growth. Boston.

Hrdlicka, A.

1905. Brain weight in vertebrates. Smithsonian Misc. Coll. **48**(1582): 89–112.

1925. Weight of the brain and of the internal organs in American monkeys. Am. J. Physical Anthrop. **8**(2): 201–211.

1930. The skeletal remains of early man. Smithsonian Misc. Coll. **83**(3033).

Jackson, C. M.

1909. On the prenatal growth of the human body and the relative growth of the various organs and parts. Am. J. Anat. **9**: 117–165.

Kappers, C. U. A.

1929. The Evolution of the Nervous System in Invertebrates, Vertebrates and Man. Bohn. Haarlem.

1936. Brain-body-weight relation in human ontogenesis and the "*Indice de valeur cérébrale*" of Anthony and Coupin. Verh. Kon. Akad. Wetenschappen Amsterdam **39**: 871–880, 1019–28.

Keith, A.

1895. Growth of brain in men and monkeys. J. Anat. Physiol. **29**: 282–303.

Klatt, B.

1918. Vergleichende metrische und morphologische Grosshirnstudien an Wild- und Haushunden. Sitz. Ber. Ges. Naturf. Fr. **2**: 35–55.

1921. Studien zum Domestikationsproblem. Untersuchungen am Hirn. Bibliotheca Genetica **2**.

Lapicque, L.

1898. Sur la relation du poids de l'encéphale au poids du corps (chez le chien). Compt. rend. Soc. Biol. Paris. **50**: 62.

1907a. Tableau du poids somatique et encéphalique dans les espèces animales. Bull. Soc. Anthrop. Paris **8**(5): 248–262.

1907b. Le poids encéphalique en fonction du poids corporal entre individus d'une même espèce. Bull. Soc. Anthrop. Paris: **8** (5): 313.

Lapicque, L., & P. Girard

1905. Poids de l'encéphale en fonction du poids du corps chez les oiseaux. Compt. rend. Acad. Sci. Paris **140**(1): 1057–1059.

1923. En fonction de la taille de l'animal le nombre des neurones sensitifs varie moins que celui des neurones moteurs. Compt. rend. Soc. Biol. **89**.

Latimer, H. B.

1938a. The prenatal growth of the cat. VII. The growth of the brain and of its parts, etc. J. Comp. Neurol. **68**: 381–394.

1938b. The weights of the brain and of its parts, of the spinal cord and of the eyeballs in the adult cat. J. Comp. Neurol. **68**: 395–404.

Lumer, H.

1936. The relations between b and k in systems of relative growth of the form $y = bx^k$. Am. Naturalist **70**: 188–191.

1940. Evolutionary allometry in the skeleton of the domesticated dog. Am. Naturalist **74**: 439–467.

Manouvrier, L.

1885. Sur l'interprétation de la quantité dans l'encéphale et dans le cerveau en particulier. Bull. Soc. Anthrop. Paris (Ser. 2) **3**(2): 137–323.

Marshall, J.

1892. On the relations between the weight of the brain and its parts and the stature and mass of the body in man. J. Anat. & Physiol. **26**: 445.

Marchand, F.

1902. Über das Hirngewicht des Menschen. Abhandl. Mathem.-Physik. Klasse Kön. Sächs. Ges. Wissensch.

Michaelis, P.

1906. Altersbestimmungen menschlicher Embryonen und Foeten auf Grund von Messungen und von Daten der Anamnese. Arch. Gynäkol. **78**: 267–288.

1907. Das Hirngewicht des Kindes. Monatsschr. f. Kinderheilk. **6**(1): 9–26.

Mollison, Th.

1910. Die Korperproportionen der Primaten. Morph. Jahrb. **42**.

1914–1915. Zur Beurteilung des Gehirnreichtums der Primaten nach ihrem Skelett. Arch. Anthrop., N. F. **13**: 388.

Pearl, R.

1905. Variation and correlation in brain weight. Biometrika **4**: 13–104.

Pfister, H.

1903. Neue Beiträge zur Kenntnis des kindlichen Hirngewichts. Archiv. f. Kinderheilk. **37**.

Rudinger, N.

1894. Über die Gehirne verschiedener Hunderassen. Sitz-Ber. Math. Phys. Abt. Akad. Wiss. München **2**: 249–255.

Scammon, R. E.

1936. Interpolation formulae for the growth of the human brain and its major parts in the first year of postnatal life. Child Develop. **7**: 149–160.

Scammon, R. E., & H. L. Dunn

1922. Empirical formulae for the postnatal growth of the brain and its major dimensions. Proc. Soc. Exp. Biol. & Med. **20**: 114–117.

Slifer, H. F.

1924. Relative brain weights in animals. Med. J. & Rec. **119**: 100.

Snell, O.

1891. Die Abhängigheit des Hirngewichts von dem Körpergewicht und den geistigen Fähigkeiten. Arch. Psychiat. **23**(2): 12.

Spitzka, E. A.

1903. Brain weights of animals, with special reference to the weight of the brain in the macaque monkey. J. Comp. Neurol. **13**: 9–17.

Warncke, P.

1908. Mitteilung neuer Gehirn- und Körpergewichtsbestimmungen. J f. Psych. & Neurol. **13**: 335–403.

Warwick, B. L.

1928. Prenatal growth of swine. J. Morph. Physiol. **46**: 59–84.

Weinert, H.

1932. Ursprung der Menschheit. Ferdinand Enke. Stuttgart.

Welcker, H., & A. Brandt

1903. Gewichtswerte der Körperorgane bei den Menschen und den Tieren. Arch. Anthrop. **28**: 1–89.

Wendt, W. W.

1909. Alte und neue Gehirnprobleme. Otto Gmelin. München.

Weygandt, W.

1928. Über Tierhirngrösse. J. f. Psych. & Neur. **37**: 394–400.

Ziehen, Th.

1901. Über vergleichend anatomische Gehirnwägungen. Monatssch. Psychiat. Neurol. **9**: 316–320.

Morphogenie des Centralnervensystems der Säugetiere. In **Hertwig**: Handbuch der vergleichenden und experimentellen Entwicklungslehre der Wirbeltiere **2**(3): 389–392.

See also, in **Bardeleben**: Handbuch der Anatomie des Menschen **4**(I,1-2): 363–381.

Zuckerman, S., & R. B. Fischer

1937. Growth of the brain in the *Rhesus* monkey. Proc. Zool. Soc. London **107**: 529–538.

TABLES 1-31

TABLE 1

BRAIN WEIGHTS AND BODY WEIGHTS OF HUMAN FETUSES

Michaelis, 1906			Various authors				Author
Age, lunar months	Av. Brain weight, gm.	Av. Body weight, gm.	Age, months	Av. Brain weight, gm.	Av. Body weight, gm.	Number (*total—185*)	
I	.. ..	2.8	4	19	138	1	G
III	4.0	17.5	4	47.3	363	3 ♀	R
IV	12.5	73.2	5	26	188	1	G
V	38.5	300.0	5	58.2	435	6 ♂	R
VI	80.4	553.6	5	58.9	428	5 ♀	R
VII	109.5	797.0	6	95.2	561	7 ♂	R
VIII	146.0	1286.0	6	87.0	623	1 ♀	R
IX	275.0	1994.0	7	108	920	1	G
			7	140.1	1061	5 ♂	R
			7	124.0	1104	7 ♀	R
			7	185.5	1175.7	30 ♂	CM
			7	184.1	1219.7	35 ♀	CM
			8	229.7	1525	2 ♂	R
			8	256.8	1764	34 ♂	CM
			8	274.81	1857	47 ♀	CM

G. Giese; R: Rüdinger; CM. Cruikshank & Miller.

TABLE 2

CHIMPANZEE

Particulars	Brain wt., gm.	Body wt., gm.	Author
♀ Premature birth	* 96	770	Coupin, *fide* A
♂ Premature birth	*270	2377	Anthony & Coupin
♂ Premature birth	*318	5490	S
♀ Premature birth	*302	5560	S
♂ 3 years	340	5500	Weber, *fide* A
♂ 3 years	347	5500	Möller, *fide* A
♂	348	6115	Weber, *fide* A
2—3 years	362	6540	Möller, *fide* A
1½—2 years	379	7430	Embleton, *fide* A
♂ 2—3 years	*412	7500	Marshall, *fide* A
4 years	381	9000	Bischoff, *fide* A
♂ 4 years	379	9400	Anthony & Coupin
4 years	367	9760	Möller, *fide* A
♀ 4 years	365	12000	A
♂ 4 years	376	14200	A
♀ 4 years	*316	16060	Anthony & Coupin
Over 4 years	391	16650	Möller, *fide* A
Over 4 years	375	19252	Owen, *fide* A
Over 4 years	376	19290	Owen, *fide* A
Over 4 years	345	21090	Meyer & Bischoff, *fide* A
♂ Over 4 years	430.5	25850	CQ
♀ Over 4 years	325	43990	CQ
♂ Over 4 years	440	56690	CQ

A: Anthony, 1928; CQ: Crile & Quiring; S: Spitska, 1903.
* Omitted in calculating the postinfantile formula.

TABLE 3

ORANGUTAN

Particulars	Brain wt., gm.	Body wt., gm.	Author
♀ complete milk dentition	*248	3170	Keith, *fide* A
♂ young	334.5	5925	Weber, *fide* A
♂ young	339	8830	Weber, *fide* A
♂ 1st molar in use	340.2	7600	Rolleston, *fide* A
♂	365	7500	Manouvrier, *fide* A
♂ young	375	11275	Weber, *fide* A
♂ 4½ years, in captivity	325.1	18600	Owen, *fide* A
♀ young	*306	20200	Weber, *fide* A
♂ adult	400	73500	Deniker & Boulart, *fide* A
♂	395	76500	Fick, *fide* A
♂	326	72575	H
♂	347.5	49895	H
♂	359	86183	H
♂	368	63503	H
♂	371	54431	H
♂	384	83915	H
♂	395	90720	H
♂	422	79379	H
♀	*267	34474	Oppenheim, *fide* H
♀	*274.5	36287	H
♀	*279	35834	H
♀	*283	36288	H
♀	*287.5	44452	H
♀	*291	34019	H
♀	*304	36741	H
♀	*322	32659	H
♀	*326	32659	H
♀	*340.5	39917	H
♀	*344	37195	H

A: Anthony, 1928; H: Hrdlicka, 1925
* Omitted in calculating the postinfantile formula Adult males were averaged, and the unweighted mean used.

TABLE 4

SEMNOPITHECUS

S. maurus, fetuses (Hulshoff Pol, *fide* Anthony, 1928)			*S. obscurus*, postnatals (Keith, 1895; also, Keith, in Anthony, 1928)		
Particulars	Brain wt., gm.	Body wt., gm.	Particulars	Brain wt., gm.	Body wt., gm.
	2	20	♀ about birth	42.8	514
	3	21	♂ complete milk dentition	62.8	2730
	9	64	♀ 1st premolar in use	57.8	2520
	12	112	♀ 2nd premolar erupting	65	3170
	21	168	♂ last canine & molar unerupted	64.4	3230
	21	172	♂ same	60.5	3630
	22	189	♂ 3 adults	67.8	6540
	24	255	♀ 7 adults	62.8	5045
	25	322			
	26	341			
	30	251			
	30	255			
New-born	32	390			
New-born	32	392			

TABLE 5

DOMESTIC CAT

Particulars	Brain wt., gm.	Body wt., gm.	Author
12-cm. fetus	* 2.3	65.8	Keith & Mies, *fide* Z
♀ new-born	* 6.5	122	Keith & Mies, *fide* Z
3 hours	* 4.8	105.6	Keith & Mies, *fide* Z
2 days	* 5.1	124	Keith & Mies, *fide* Z
4 days	* 6.7	162.7	Keith & Mies, *fide* Z
8 days	* 8	206.3	Keith & Mies, *fide* Z
♂ 10 days	* 9	207	A
♂ 10 days	* 9	222	A
♂ "young"	*16.1	338	Weber, *fide* A
♀	*21	490	A
♀ 2 months ca.	21	600	A
♀ 2½ months	22.5	701	Weber, *fide* A
♂ "young"	26	990	A
♀ 3½ months	26.5	1220	Weber, *fide* A
♂ "young"	24	1442	A
♂	28	1888	A
♂ "young"	27	2350	A
♂ 6 months ca.	30	2500	A
♂ 7 months	28	3045	A
♂ 5 adults	*31	3300	Weber, *fide* A

A: Anthony, 1928; Z. Ziehen, 1906.
* Not used in the calculations.

TABLE 6

SHEEP

Fetuses (Ziehen, 1906)		Postnatals (Crile & Quiring, 1940)		
Brain wt., gm.	Body wt., gm.	Brain wt., gm.	Body wt., gm.	Number
2.5	57.5	*73.6	6870	8
3.61	60.0	*71.5	8550	5
3.36	68.3	88.6	15760	4
3.75	68.0	109	40230	9
4.43	98.0	106.5	52100	7
5.07	127.2			
5.34	114.7			
5.23	131.7			
5.79	121.0			
5.84	154.0			
8.33	182.3			
8.15	173.5			
7.65	208			
8.67	242.4			
10.02	273.5			
12.01	369			
15.2	480			
19.8	666			
17.35	571			

* Not used in calculating the postinfantile formula.

TABLE 7

Ox

Fetuses (Ziehen, 1906)		Postnatals (Crile & Quiring, 1940; steers omitted)		
Brain wt., gm.	Body wt., gm.	Brain wt., gm.	Body wt., gm.	Number
1.72	28.55	*193.9	19500	1
2.37	43.85	*215.2	22000	1
3.21	55.9	304	51900	3 ♂
4.83	91.4	299	90000	6 ♂
5.72	159	356	214000	5 ♂
6.91	174.5	386	241000	3 ♂
9.23	349	384	367000	2 ♂
11.94	384	*357	371000	7 ♀
9.72	393	*408	412000	213 ♂
15.31	559	*408	413000	218 ♀
16.45	746	*403	450000	62 ♀
		*473.5	489900	1 ♀
		*417	491000	44 ♀
		*420	506000	71 ♀
		*408	552000	5 ♂
		*415	574000	200 ♀
		*444	591000	2 ♂
		*447	597000	10 ♂
		*471	861000	2 ♂
		*462	888000	5 ♂

* Omitted from the calculation of the postinfantile formula.

TABLE 8

FETAL PIG

(*Pooled data of Anthony, 1928, and Ziehen, 1906. Anthony's source: L G. Lowry.*)

Number	Brain wt., gm.	Body wt., gm.	Author
3	.02	.29	A
5	.06	.70	A
6	.10	1.70	A
5	.19	3.25	A
2	.34	4.97	A
6	.69	10.33	A
4	1.37	28.20	A
1	1.11	42.50	Z
1	1.68	45.00	Z
1	2.35	56.10	Z
10	2.00	63.54	A
1	2.55	72.60	Z
4	2.47	74.90	A
1	2.54	77.50	Z
1	2.45	76.00	Z
1	2.45	76.50	Z
1	2.61	77.50	Z
1	2.35	78.00	Z
8	3.20	90.20	A
3	3.86	97.00	A
1	2.50	100.00	Z
1	2.60	105.50	Z
22	5.10	126.45	A
3	6.22	153.00	A
6	6.88	216.50	A
1	8.13	283.00	Z
5	10.25	288.80	A
5	11.01	334.70	A
5	13.78	395.00	A
3	18.97	465.00	A
1	13.83	485.00	Z
1	9.26	535.00	Z
5	25.21	731.00	A
3	29.50	745.00	A
1	33.04	826.00	A
1	30.85	858.00	Z
1	34.45	1352.00	Z

TABLE 9

HORSE

(Crile & Quiring, 1940; geldings omitted.)

Number	Average age	Brain wt., gm.	Body wt., gm.
	Fetuses		
3 ♀	91 days, premature	183.3	13000
1 ♀	71 days, premature	226	19500
5 ♀	46.4 days, premature	254.4	26470
5 ♂	54.4 days, premature	273.9	27300
1 ♀	60 days, premature	ca. 242	31750
11 ♀	14.6 days, premature	333.9	47680
15 ♂	16 days, premature	317.3	38910
	Infants		
19 ♀	5.6 days old	366.5	54320
18 ♂	3.1 days old	370.1	52450
1 ♂	3 months old	400	92980
3 ♂	33.5 days old	425.3	93890
4 ♀	83 days old	470.2	116770
1 ♀	3 months old	475	118800
	Postinfantile, but immature (only this category used in calculating formula)		
1 ♀	6-7 months old	492	181400
2 ♀	1 year old	525	184160
8 ♂	9 months old	582.8	285130
2 ♂	1 year old	588	300500
5 ♂	1 year old	602.4	306350
2 ♀	1 year old	616	380110
7 ♀	2-3 years old	632	408500
3 ♂	2-3 years old	621.4	433920
	Adults		
21 ♂ ♀		520	230900
31 ♀	14 years	468	262000
1 ♂	30 years	573	362800
1 ♂		504	362870
1 ♀	18 years	604	376480
1 ♀	25 years	692	380750
1 ♀	15 years	690	402650
10 ♀	19.1 years	637.7	443360
1 ♂	27 years	618	461760
5 ♂	17.2 years	706.7	485310
1 ♀	12 years	655	521640

TABLE 10

PROSPECTUS OF THE FORMULAE FOR POSTINFANTILE AND FOR FETAL PERIODS

Genus		Postinfantile	Fetal
1. *Homo*	♂	X	X
	♀	X	
2. *Pan* ♂ ♀		X	
3. *Simia* ♂		X	
4. *Rhesus*	♂	X	
	♀	X	
5. *Semnopithecus* ♂ ♀		X	X
6. *Felis* ♂ ♀		X	
7. *Mus*	♂	X	
	♀	X	
8. *Ovis* ♂ ♀		X	X
9. *Bos*	♂	X	
	♂ ♀		X
	♀		
10. *Sus* ♂ ♀			X
11. *Equus* ♂ ♀		X	

TABLE 11

GROWTH OF BRAIN WEIGHT/BODY WEIGHT AFTER INFANCY

Genus		Number	$\log y = \log a + b \log x$	$y = ax^b$
1. *Homo*	♂	89	2 8431 + 06326 log x	696 $8x^{06326}$
	♀	62	2 7711 + 07010 log x	590 $37x^{07010}$
2. *Pan* ♂ ♀		17	2 408 + 0389 log x	255 $9x^{0389}$
3. *Simia* ♂		16	2 4204 + 0322 log x	263 $3x^{0322}$
4. *Rhesus*	♂	36	1 2346 + 19626 log x	17 $163x^{19626}$
	♀	40	1 384 + 1449 log x	24 $21x^{1449}$
5. *Semnopithecus* ♂ ♀		16	1 4316 + 1019 log x	27 $01x^{1019}$
6. *Felis dom.* ♂ ♀		10	8331 + 1835 log x	6 $809x^{1835}$
7. *Mus norv.*	♂	211	− 0998 + 1838 log x	$7944x^{1838}$
	♀	261	− 1231 + 1926 log x	$7533x^{1926}$
8. *Ovis ar.* ♂ ♀		20	1 1505 + 189 log x	14 $14x^{189}$
9 *Bos taur.* ♂		19	1 5678 + 1868 log x	36 $966x^{1868}$
11. *Equus cab.* ♂ ♀		30	1 672 + 1997 log x	46 $98x^{1997}$

TABLE 12

PRENATAL GROWTH OF BRAIN WEIGHT/BODY WEIGHT

	Number	$\log y = \log a + b \log x$	$y = ax^b$	Body wt., gm. (*Approx.*,	Brain wt., gm. *at birth*)	Ratio of brain wt./ body wt.
1. *Homo* ♂ ♀ {Composite data Michaelis' means	185 ?	$-$ 86306 + 1 0098 $\log x$ $-$ 606 + 914 $\log x$	$.13707x^{1.0098}$ $.2477x^{914}$	3200	351*	.11
5. *Semnopithecus* ♂ ♀	14	$-$ .557 + .8075 $\log x$	$2773x^{8075}$	391**	32**	.08
8. *Ovis* ♂ ♀	19	$-$ 885 + .774 $\log x$	$1300x^{774}$	3400***	52***	.015
9. *Bos* ♂	11	$-$ 661 + 655 $\log x$	$2183x^{.655}$	20000	205	010
10. *Sus* ♂ ♀	130	-1 22885 + 9112 $\log x$	$05904x^{.9112}$	1450 826***	32 34***	.022 .041

* Boyd, 1942; ** Hulshoff Pol, *fide* Anthony, 1928; *** Harris, 1929.

TABLE 13

ADULT UNGULATES

(*Table 1 from Brummelkamp, 1939c; adapted.*)

Entry No.	Genus and species	Brain wt., gm.	Body wt., gm.	Ceph. coeff.
1	*Elephas indicus*	4717	3048000	1 184
2	*Elephas indicus*	4048	2047000	1 234
3	*Camelus dromedarius*	762	400000	.589
4	*Giraffa camelopardalis*	680	529000	.450
5	*Equus caballus*	532	368000	.431
6	*Alces americanus*	407	272000	390
7	*Equus asinus*	385	175000	470
8	*Oryx beisa*	280	107000	450
9	*Antilope caāma*	269	99500	.451
10	*Antilocapra americanus*	130 2	34474	393
11	*Antilope cervicapra*	90	13500	457
12	*Gazella*	216	68000	446
13	*Gazella isabella*	81 6	12170	440
14	*Ovis tragelaphus*	209	56000	.481
15	*Ovis aries*	140	55000	326
16	*Ovis aries*	94.5	28000	320
17	*Sus scrofa* (*wild*)	178	56000	410
18	*Sus scrofa* Berkshire	125	104000	204
19	*Sus scrofa* Middlewhite	125	104000	204
20	*Sus scrofa* ("Edelschwein" ♀)	168	281000	158
21	*Sus scrofa domestica*	113	150000	151
22	*Sus scrofa domestica*	90 94	104000	148
23	*Cervus elaphus*	411	125530	605
24	*Cervus dama*	224 5	37195	649
25	*Cervus muntjak*	125	16600	566
26	*Cervus porcinus*	142	30000	462
27	*Cervus hippelaphus*	229	73500	454
28	*Cervus capreolus*	97 5	14500	475
29	*Rupicapra rupicapra*	118 5	26500	.414
30	*Cephalophus maxwelli*	41 1	3780	.424
31	*Cephalophus maxwelli*	38	3357	.417
32	*Cephalophus maxwelli*	37 5	3130	.429
33	*Cephalophus maxwelli*	35 4	3160	403
34	*Damaliscus lunatus*	324	82000	603
35	*Hippopotamus amphibius*	582	1755000	.198
36	*Bos taurus* Monthéliard	493	660000	288
37	*Bos taurus* Vendée	480	540000	.314
38	*Bos taurus*	423	465000	.301
39	*Tapirus indicus*	265	201000	.300
40	*Tapirus americanus*	169	160000	.217
41	*Capreolus caprea*	103 5	15422	489
42	*Capreolus caprea*	98	15000	.469
43	*Capreolus caprea*	93	14062	461
44	*Phacochoerus africanus*	132 5	67000	276
45	*Tayassu tajac*	101	19618	416
46	*Tragulus napu*	18 3	2670	228
47	*Tragulus memmina*	17 1	2368	229
48	*Tragulus javanicus*	15.85	2037	230
49	*Mesohippus bairdi*	88	27000	303
50	*Mesohippus bairdi*	63	15000	301
51	*Palaeosypos leidyi*	102	380000	081
52	*Anoplotherium commune*	71	200000	080
53	*Moeritherium*	173	336000	.147
54	*Diplobune bavaricum*	47.5	38500	135
55	*Uintatherium mirabile*	150	2000000	.047
56	*Coryphodon hamatum*	44	235000	.045

TABLE 14

ADULT RODENTS

(Table 1 from Brummelkamp, 1939b; adapted.)

Entry No.	Genus and species	Brain wt., gm.	Body wt., gm.	Ceph. coeff.
1	*Hydrochaerus capybara*	75	28500	251
2	*Dasyprocta agouti*	20	2684	249
3	*Sciurus rufiventer*	9 2	650	.252
4	*Sciurus rufiventer*	8.95	580	261
5	*Sciurus carolinensis*	7 58	469	249
6	*Sciurus carolinensis*	7 48	466	246
7	*Sciurus vulgaris*	6 10	323	246
8	*Sciurus vulgaris*	5 81	287	.250
9	*Sciurus hudsonicus*	4 103	159	240
10	*Hystrix*	37 5	15000	179
11	*Lepus timidus*	16 7	3833	171
12	*Lepus cuniculus* dom. gem.	11 20	3375	123
13	*Lepus cuniculus* (average)	9 3	1226	179
14	*Oryctolagus cuniculus ferus*	10 4	1440	183
15	*Pteromys nitidus*	11 8	1600	196
16	*Cavia cobaya*	4 73	700	124
17	*Cavia porcellus*	4 54	675	122
18	*Dipus hirtipes*	1 85	73	171
19	*Sciuropterus volans*	1 92	64	190
20	*Lagostomus trichodactylus*	8 8	3854	089
21	*Arvicola agrestis*	9	42 5	112
22	*Arvicola terrestris*	355	90 25	027
23	*Mus agrarius*	20	33 3	028
24	*Mus norvegicus*	2 36	448	077
25	*Mus rattus*	1 59	200	084
26	*Mus musculus*	43	20.85	078
27	*Mus musculus*	36	16	077
28	*Mus sylvaticus*	59	21 6	107
29	"Japanese mouse"	.35	7 85	111
30	*Cricetulus griseus*	.628	23 18	109
31	*Mus musculus*	475	24.32	080
32	*Mus wagneri*	424	18.92	082

TABLE 15

ADULT CARNIVORES

Entry No.	Genus and species	Number	Brain wt., gm.	Body wt., gm.	Author
			Viverridae, Hyenidae		
1	*Genetta tigrina*	3 ♂ ♀	15 71	1376 3	CQ
2	*Ichneumia albic.*	1 ♂	28 30	4400	CQ
3	*Crocuta croc.*	2	175	62370	CQ
			Felidae		
4	*Felis domest.*	10 ♂ ♀	25.30	3275 6	CQ
5	*F. leo*	5 ♂	240 72	150720	CQ
6	*F. pardus*	1 ♂	135	48000	CQ
7	*F. oregonensis*	1 ♂	106 7	28790	CQ
8	*F. bangsi*	1 ♂	129	25960	CQ
9	*F. onca*	1 ♀	147	34470	CQ
10	*F. capnesis*	3 ♂ ♀	57 687	7197 7	CQ
11	*F. tigris*	2 ♂ ♀	263 5	184500	CQ
12	*F. ocreata*	1 ♀	28 48	2700	CQ
32	*F. pardalis*	2 ♂	63 1	9525 5	H
33	*F. concolor*	1 ♂	154	54432	H
34	*F. cacomistli*	1 ♀	41 9	2722	H
35	*F. serval*	1 ♂	54 1	11340	H
36	*Lynx canad.*	1 ♂	69 5	14969	H
37	*Lynx rufus*	1 ♂	65 0	6350	H
			Procyonidae		
14	*Potos flavus*	1 ♀	31 05	2620	CQ
15	*Procyon lotor*	4 ♂ ♀	39 19	4288 25	CQ
38	*Nasua rufa*	1 ♀	34 0	3175	H
39	*Potos caudiv*	1 ♀	35 1	1863	H
			Canidae		
16	*Vulpes lagopus*	1 ♂	44 50	3385	CQ
17	*Vulpes fulva*	1 ♀	53 30	4625	CQ
18	*Urocyon ciner.*	1 ♂	37 28	3749	CQ
19	*Otocyon megal*	1 ♀	26 09	3335	CQ
20	*Thos mesomelas*	2 ♂	46 00	2850	CQ
21	*Canis latrans*	1 ♀	84 24	8510	CQ
22	*C. familiaris*	9 ♂ ♀	78 945	13404 4	CQ
23	*C. familiaris*	1 ♂	105 9	24490	CQ
24	*C. familiaris*	3 ♂	113 57	30146 7	CQ
25	*C. familiaris*	6 ♂ ♀	87 287	25423 3	CQ
26	*C. lupus*	1 ♂	119	22680	CQ
27	*C. lubilis*	1 ♂	152	29940	CQ
			Ursidae		
28	*Ursus horribilis*	1 ♀	233 9	142880	CQ
29	*Thalarctos mar.*	2 ♂ ♀	498	258285	CQ
40	*Ursus horribilis*	1 ♂	389	149688	H
41	*Ursus torquatus*	1 ♂	269	69860	H
42	*Melursus urs.*	1 ♀	267	136080	H
43	*Helarctor malay.*	1 ♀	385 5	45020	H
			Mustelidae		
30	*Mustela arctica*	4 ♂ ♀	5 0975	157 22	CQ
31	*Mephitis meph.*	3 ♂ ♀	10 10	2073 3	CQ
44	*Putorius putor*	1 ♂	7 87	915	H

CQ: Crile & Quiring; H: Hrdlicka.

TABLE 16

ADULT PRIMATES

Entry No.	Genus and species	Number	Brain wt., gm.	Body wt., gm.	Author
2	*Semnopithecus entellus*	1 ♂	117	6647	S
3	*Cercopith. sp.*	4 ♂ ♀	59 75	2469	S
4	*Cercopith. mona*	1 ♂	67	3001	S
5	*Cercopith. griseo-vir.*	1 ♂	72	3202	S
6	*Chlorocebus sab.*	1	71	1480	S
7	*Chlorocebus cynoc.*	1	68 5	3880	S
8	*Cercocebus fulig.*	2 ♂ ♀	98	1375 5	S
9	*Macacus rhes.*	24 ♂	79 94	1598 46	S
10	*Macacus rhes.*	24 ♀	78 52	1941.21	S
11	*Macacus rhes.*	7	88 57	1783 57	S
12	*Macacus cynomolg.*	4 ♂	62 5	1653 75	S
13	*Macacus cynomolg.*	5 ♀	61 2	1354.2	S
14	*Macacus nemestr.*	5 ♂	118 4	6786	S
15	*Macacus nemestr.*	3 ♀	99 33	2989	S
16	*Macacus sinus*	3 ♂	67	963 67	S
17	*Macacus pileatus*	5 ♂ ♀	63 1	1402 6	S
18	*Macacus speciosus*	1 ♂	98	5560	S
19	*Macacus melanotus*	1 ♂	80	1105	S
20	*Cynopithecus niger*	2 ♂	108	5920	S
21	*Cynocephalus bab.*	5 ♂ ♀	137 4	2541 8	S
22	*Cynocephalus hamad.*	1 ♂	199	10230	S
23	*Cynocephalus hamad.*	6 ♀	121 17	2224 17	S
24	*Cynocephalus anubis*	1 ♀	152	5000	S
25	*Cynocephalus sphinx*	1 ♀	135	2481	S
26	*Cynocephalus leucoph.*	1 ♀	124	1718	S
27	*Mycetes cavaya*	1 ♂	45	826	S
28	*Mycetes ursinus*	1 ♀	19	1025*	S
29	*Ateles beelzebus*	1 ♀	97 5	1870	S
30	*Lagothrix humb.*	1	112	4850	S
31	*Cebus capuc.*	5 ♂ ♀	70 2	1453 4	S
32	*Cebus capill.*	2 (♂)	73	1515	S
33	*Cebus hypoleuc.*	3 ♂	54 33	587	S
34	*Cebus subcrist.*	1 ♂	71	902	S
35	*Cebus albifrons*	1 ♂	58	675	S
36	*Hapale penic.*	1 ♀	8	206	S
37	*Iacchus vulg.*	8 ♂ ♀	7 81	219.75	S
38	*Midas urs.*	1 ♀	24	361	S
39	*Lemur brun.*	1 ♂	26	1505	S
40	*Nycticebus tard.*	1 ♂	12	612	S
41	*Simia satyrus*	8 ♂	371.6	72575	H
42	*Simia satyrus*	11 ♀	299.8	36439	H
43	*Hylobates sp.*	3 ♂ ♀	101 7	6125 7	H
44	*Hylobates agil.*	9 ♂	87 45	5946	H
45	*Hylobates agil.*	3 ♀	79 8	5180	H
46	*Hylobates mulleri*	3 ♂	96.8	5516	H
47	*Hylobates mulleri*	4 ♀	92 6	6348	H
48	*Symphalangus*	3 ♂ ♀	133 5	12079	H
49	*Oedipomidas*	3 ♂ ♀	9.35	305.2	H
50	*Homo*	41 ♂ ♀	1320 15	62080	CQ
51	*Leontocebus geoffroyi*	184 ♂ ♀	8.21	275.5	CQ

S: Spitzka, 1903; H: Hrdlicka, 1925; CQ: Crile & Quiring, 1940.
* Misprint?

TABLE 17

ADULTS OF OTHER EUTHERIAN ORDERS

Genus and species	Number	Brain wt., gm.	Body wt., gm	Author
		Edentates		
Tamanduas tetradactyla	2 ♂ ♀	25	3692	CQ
Dasypus novemcinctus	10 ♂ ♀	7 5	3401	CQ
Bradypus griseus	12 ♂ ♀	15 33	3156 5	CQ
Choloepus hoffmanni	8 ♂ ♀	23 44	4880.75	CQ
		Pinnipedia		
Erignathus barabtus	1 ♀	460	281000	CQ
Odobenus rosmarus	1 ♂	1126	667000	CQ
Phoca richardi	1 ♂	442	107300	CQ
Phoca hispida	5 ♂ ♀	253	39570	CQ
Phoca vitulina	5 ♂ ♀	270 5	12610	H
		Chiroptera		
Desmodus rotundus	5 ♂ ♀	8 95	818 2	CQ
		Insectivora		
Scalopus aquaticus	1 ♂	1 16	39 6	CQ
Blarina brevicauda	68 ♂ ♀	347	17 36	CQ
		Cetacea		
Phocaena phocaena	1 ♂	1735	142130	CQ
Delphinapterus leucas	6 ♂ ♀	2351 7	395280	CQ
Balaenoptera musculus	1	6800	58059000	CQ

CQ Crile & Quiring, 1940, H Hrdlicka, 1925

TABLE 18

CRANIAL CAPACITIES OF SOME PALEANTHROPOI, AND CALCULATIONS OF BRAIN WEIGHTS

(*Hrdlicka, 1930.*)

Specimen	Volume measured by	cc	Brain wt., gm (*mihi*)
Ehringsdorf		1450?	1270?
Eoanthropus	Smith Woodward	1300	1139
	Elliot Smith	1200	1051
	Keith	1400	1226
Gibraltar	Sollas	1250	1095
	Keith	1200	1051
La Chapelle	Boule	1600–1626	1401–1424
La Quina	H. Martin	1350	1182
Le Moustier	Weinert	1564	1370
Neanderthal	Schaffhausen	1033	
	Huxley	1230	1077
	Schwalbe	1234	1081
Pithecanthropus	Dubois	1000+	
	Dubois	900	789*
	Weinert	1000	870–920†
Rhodesia	Elliot Smith	1280	1120
Rome	Sergi	not over 1200	ca. 1050

* Using Bolk's ranges and 900 cc., one gets 664–846 gm.
† Weinert's estimate.

TABLE 19

THE *Homo* LINE

$(Y = -2\ 2417 + 1\ 54915\ X - .0899\ X^2)$

Groups	Empirical X	Empirical Y	Calculated Y	Source of raw data see
68 Blarinae	1 2396	− 4597	− 4597	TABLE 17
184 Hapalidae	2 4401	9143	1 0028	TABLE 16, item 51
15 Cebidae	3 1498	1 8422	1 7458	TABLE 16, items 27, 29–35
* 93 Cercopithecidae	3 3021	1 8949	1 8928	TABLE 16, items 2–19
17 Baboons	3 4688	2 1057	2 0493	TABLE 16, items 20–26
* 41 Homines	4 7929	3 1206	3 1273	TABLE 16, item 50

* Data used for calculating the line

TABLE 20

RELATIVE INCREASES OF CEPHALIZATION IN SOME PRIMATE STAGES, ASSUMING THE INSECTIVORES AS A BASE

Groups compared	Differences of Y	Differences of X	Ratio
Marmosets / Shrews	1 495	1 2005	1 245
Cercopithecidae / Marmosets	8596	8620	996
Baboons / Cercopithecidae	2108	1667	1 265
Humans / Baboons	1 0149	1 3241	767
Humans / Cercopithecidae	1 2257	1 4908	822

TABLE 21

BRAIN AND BODY WEIGHTS, AND CEPHALIZATION COEFFICIENTS, OF SOME REPTILES

Entry No	Number	Genus and species	Body wt, gm	Brain wt, gm	Ceph coeff
		*Crile & Quiring's Data, 1940**			
R 1	1 ♂	*Alligator miss*	109000	8 40	012715
	1 ♂	*Alligator miss*	52400	7 23	01656
	1 ♂	*Alligator miss*	173000	11 20	01289
	1 ♂	*Alligator miss*	205000	14 08	01498
	1 ♂	*Crocodilus amer*	134000	15 60	02109
R 2	1 ♀	Gila Monster	514	729	02219
R 3	1 ♀	Iguana	4190	1 44	01353
R 4	15 ♂ ♀	*Lacerta viridis*	50	121	01356
R 5	3 ♂ ♀	Black Snakes	431	291	00977
R 6	20 ♂ ♀	*Tropidonotus*	70	100	00927
R 7	6 ♂ ♀	*Zamenis*	220	209	01022
R 8	30 ♂ ♀	*Emis*	250	25	01141
R 9	30 ♂ ♀	*Testudo*	320	30	01189
R 1		*Mean for Crocodilia*			01565
		Mean for Ophidia			00952
		Median of reptilian means			01258

Entry No	Genus and species	Body wt gm	Brain wt, gm	Ceph coeff
	Dubois' Data, 1913†			
R 10	Monitor	7500	2 44	0165
R 11	Little Gecko	4 7	043	0181
R 12	Emerald Lizard	16 8	093	0191
R 13	Cobra	1770	646	0098
R 14	Common Viper	64 2	105	0102
R 15	Common Lizard	12 507	076	0185
R 16	Slow Worm	16 252	039	0082
	Slow Worm	18 9	037	0071
R 17	Greek Tortoise	993 58	360	0075

* Cephalization coefficients mine
† Cephalization coefficients his

TABLE 22

BRAIN AND BODY WEIGHTS, AND CEPHALIZATION COEFFICIENTS, OF SOME AMPHIBIA

(*Table 1 from Brummelkamp, 1939d, adapted; his sources: Dubois, Donaldson.*)

Entry No.	Genus and species	Brain wt., gm.	Body wt., gm.	Ceph. coeff.
A 1	*Rana catesbyana*	204	244 40	009614
A 2	*Rana virescens*	153	73.35	.01407
A 3	*Rana esculenta*	.1137	52.90	.01254
A 4	*Rana esculenta*	.106	44 50	.01287
A 5	*Rana esculenta*	.1007	39 21	01312
A 6	*Rana pipiens*	1165	44.20	01419
A 7	*Rana temporaria*	.0909	34.80	.01265
A 8	*Rana temporaria*	0861	31.30	.01272
A 9	*Rana fusca*	088	53 00	009696
A 10	*Hyla arborea*	.043	4 80	01799
A 11	*Bufo vulgaris*	.073	44 50	008862
A 12	*Alytes obstetricus*	.041	7.70	01319

TABLE 23

HUMAN BRAIN WEIGHED BY PARTS; RATIO BETWEEN PONS-MEDULLA AND ENCEPHALON

(*Data of Marshall, 1892.*)

I umber	II Ages	III Av. body wt., lbs.	IV Av. encephalon wt., ozs.	V Pons-medulla wt., ozs.	VI Ratio of V/IV
			Males		
55	20–30	92.14	47.9	.93	.0194
103	30–40	93 35	48.2	.98	.02012
135	40–50	102	47.75	1.06	0222
110	50–60	102 5	47.44	.98	.02065
123	60–70	103 13	46.16	97	0210
102	70–80	106.13	45.5	94	02064
24	80–90	99	45.34	89	01961
Totals and means 652	20–90	99 75	46 88	97	0207
			Females		
70	20–30	86.13	43.7	.88	.0201
85	30–40	87	43.09	.91	.0211
97	40–50	84	42.81	.89	02075
100	50–60	86	43 12	.86	.01992
142	60–70	86.14	42.69	.83	.0194
146	70–80	80.4	41.27	.88	.0213
75	80–	79.5	39.77	.82	.0206
Totals and means 715	20–	84.19	42.35	87	.0205

TABLE 24

PROSPECTUS FOR TABLE 25

(*Orders of mammals, number of genera listed, and operators for weighting the genera*)

Order	Number of genera	Weighting operator
Primates	23	1 304
Carnivora	23	1 304
Ungulata	22	1 364
Rodentia	30	1 000
Edentata	4	7 500
Pinnipedia	3	10 000
Chiroptera	4	7 500
Insectivora	5	6 000
Cetacea	2	15 000

TABLE 25

GENERA FROM PRECEDING TABLES, GROUPED ACCORDING TO LOG BODY WEIGHT

(*Some genera have been split because of wide species range in log body weight*)

Group	log body wt not more than	Genera	Mean log brain wt	Mean log body wt
I	1 50	*Mus, Desmodus, Rhinolophus, Vespertilio, Blarina*	− 3533	1 345
II	2 00	*Dipus, Sciuropterus, Arvicola, Mus, Tupaya, Talpa, Scalopus*	1291	1 786
III	2 50	*Iacchus Hapale, Oedipomidas, Mustela, Sciurus, Mus, Geomys*	714	2 304
IV	3 00	*Mycetes, Midas, Nyctocebus, Putorius, Sciurus, Cavia, Mus, Perodipus, Pteropus, Erinaceus*	875	2 851
V	3 50	*Cercopithecus, Cercocebus, Macacus, Cynocephalus, Ateles, Cebus, Lemur, Genetta, Potos, Thos, Mephitis, Tragulus, Dasyprocta, Lepus, Oryctolagus, Pteromys, Fiber, Bradypus*	1.43	3 326
VI	4 00	*Semnopithecus, Cynopithecus, Lagothrix, Hylobates, Ichneumia, Felis, Procyon, Vulpes, Urocyon, Otocyon, Nasua, Cephalophus, Lepus, Lagostomus, Capromys, Tamanduas, Dasypus, Choloepus*	1.437	3.740
VII	4 50	*Symphalangus, Felis, Canis, Antilope, Rupicapra, Capreolus, Tayassu, Felis, Hystrix, Hydrochaerus*	1 98	4 282
VIII	5 00	*Simia, Gorilla, Pan, Homo, Hyaena, Felis, Helarctor, Antilope, Antilocapra, Gazella, Ovis, Sus, Cervus, Damaliscus, Phoca*	2 41	4.69
IX	5 50	*Ursus, Thalarctos, Melursus, Equus, Alces, Oryx, Tapirus, Erignathus, Phocaena*	2 855	5 26
X	6 00	*Camelas, Giraffa, Odobenus, Delphinapterus*	3 218	5 70
XI	6 50	*Loxodontia, Hippopotamus*	3 208	6 32

TABLE 26

EXPONENTS OF CEPHALIZATION

$(Y = A + bX + cX^2.)$

Group	Exponent		
1. Reptiles	−1 7095	+ .3679 X	+ 03613 X^2
2. Marsupials, all genera	−1 2332	+ 6993 X	−.00492 X^2
3. Marsupials, all genera plus ($Y = -1.900$, $X = 0$)	−1.9140	+1 1194 X	− 0679 X^2
4. Mammals, von Bonin's data	−1.4045	+ 8742 X	− 01363 X^2
5. Mammals, data of TABLE 25	−1 8080	+1 1540 X	− 0543 X^2
6. 10 mammalian "stocks" (a)	−1 9510	+1 3262 X	− 0871 X^2
7. 10 mammalian "stocks" (b)	−1.9301	+1.3222 X	− 0883 X^2
8. Rodents[3]	−1.0146	+ .6268 X	− 00141 X^2
9. Sciuridae[1,4]	−1 3835	+ 9350 X	− 0449 X^2
10. Chiroptera[2,3,4]	−1 8997	+1 2603 X	−.0971 X^2
11. Edentates[2,3,4]	−1.9001	+1.3002 X	−.1112 X^2
12. Proboscidea[1,3]	−1 8069	+1.1426 X	− 0475 X^2
13. Elephants[1,2,4]	−1 9005	+1 2358 X	−.0591 X^2
14. Perissodactyla[1,3]	−2 3085	+1.6561 X	− 1333 X^2
15. Perissodactyla[1,2,4]	−2.0840	+1.6339 X	− 1367 X^2
16. Pecora (Antelopinae)[1,3,4]	−1.9284	+1.2890 X	− 0838 X^2
17. Anoplotheria[2]	−1.9033	+1 2442 X	− 1012 X^2
18. Amblypoda[1,2]	−1 9180	+1 1818 X	− 0902 X^2
19. Carnivora[1,3,4]	−1.9227	+1 2830 X	−.0844 X^2
20. Pinnipedia[1,3,4]	−2 1403	+1 4872 X	− 1064 X^2
21. Cetacea[1,3]	−2.3473	+1 6680 X	− 1162 X^2
22. Large whales[1,2,4]	−1 9000	+1 2487 X	− 0694 X^2
23. Primates (*Homo*)[1,3,4]	−2 2417	+1 5491 X	−.0899 X^2

[1] Line assumed as passing through *Blarina*.
[2] Line assumed as passing through reptilian value $Y = -1.900$, $X = 0$
[3] Used for calculating line of 10 mammalian "stocks" (a), No. 6, above.
[4] Used for calculating line of 10 mammalian "stocks" (b), No. 7, above.

TABLE 27

EXPONENTS OF CEPHALIZATION

($Y' = a' + b'X' + c'X'^2$, *where the origin of the coordinate axes is taken at* $X = .2385$, $Y = -1.6217$. *Entry numbers the same as in* TABLE 26.)

Group	Exponent
1. Reptilia	.3881 X' + .03546 X'^2
7. Mammalia (modern)	1.2799 X' − 08841 X'^2
9. Sciuridae	1.1730 X' − 07942 X'^2
16. Antelopinae (Pecora)	1.2474 X' − .083916 X'^2
19. Carnivora	1.2442 X' − .084915 X'^2
20. Pinnipedia	1.3389 X' − 093466 X'^2
23. *Homo*	1.3572 X' − 06993 X'^2

TABLE 28

A CHECK ON THE *Homo* FORMULA OF TABLE 27

Group	A	B	C	D	E	F
Blarina	1 2396	1.0011	– 4597	1 1620	1 2875	1 1620
Hapale	2 4401	2 2016	1 0028	2 6245	2 6503	2 6570
Cercopithecidae	3 3021	3 0636	1 8928	3 5145	3 5018	3 5166
Baboons	3 4688	3 2303	2 0493	3 6710	3.6550	3 7274
Homo	4 7929	4 5644	3 1273	4 7490	4 7420	4.7423

A values of X from TABLE 19
B $X' = X - 2385$
C values of Y calculated from TABLE 26
D $Y' = Y + 1\ 6217$
E Y' calculated from TABLE 27
F empirical Y from TABLE 19, plus 1 6217

TABLE 29

THE VERTICAL DISTANCE, RY, BETWEEN A MAMMALIAN PARABOLA, $b'X' - c'X'^2$, AND A REPTILIAN, $b'X' + c'X'^2$

(*Entry numbers those of* TABLE 26.)

Group	$RY = mX' - nX'^2$.
7 Mammalia	$8918\ X' - 12387\ X'^2$
9. Sciuridae	$7849\ X' - 11484\ X'^2$
16. Antelopinae	$8593\ X' - 11938\ X'^2$
19. Carnivora	$8561\ X' - 12038\ X'^2$
20. Pinnipedia	$9508\ X' - 12893\ X'^2$
23. *Homo*	$9691\ X' - 10539\ X'^2$

Table 30

Values of log Body Weight X, where the Difference between the Mammalian log Brain Weight and that of a Reptile of Equivalent Body Weight is a Maximum

(Column A calculated from Table 26; *Column B, from* Table 27. *Entry numbers those of* Table 26.)

Group	A	B
7:1 Mammalia	3.83	3 838
9:1 Sciuridae	3 49	3 648
16:1 Antelopinae	3.84	3.838
19:1 Carnivora	3.80−	3.788
20:1 Pinnipedia	3.93	3.928
23:1 *Homo*	4 69	4.828

Table 31

Values of l, where $l = \dfrac{\lambda(\text{mammal})}{\lambda(\text{reptile})}$; λ Being the Slope of any Parabola in Table 27

(Entry numbers those of Table 26.)

7:1 Mammalia	$3.2542 - .8825\,X' + .06872\,X'^2$
9:1 Sciuridae	$3.0865 - 8010\,X' + .06212\,X'^2$
16:1 Antelopinae	$3.16897 - 8475\,X' + 06575\,X'^2$
19:1 Carnivora	$3.1653 - .8541\,X' + .06653\,X'^2$
20:1 Pinnipedia	$3.7547 - 1\ 1490\,X' + .10387\,X'^2$
23:1 *Homo*	$3.4492 - .8126\,X' + .05849\,X'^2$

FIGURES 1-24

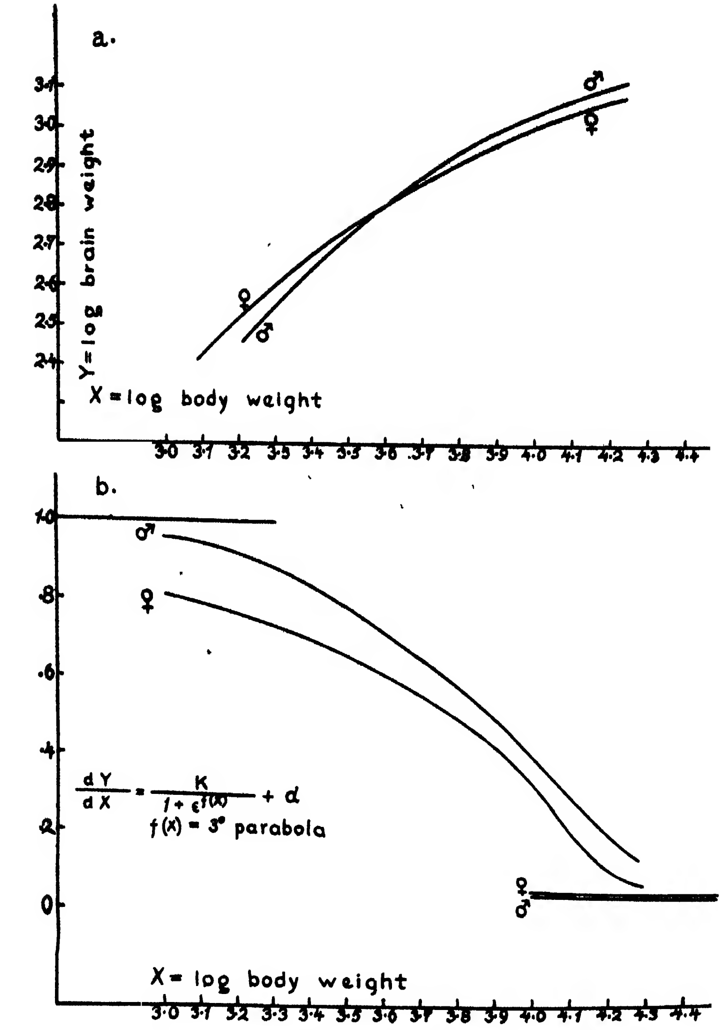


FIGURE 3. The infantile period. Detail of FIGURES 1 and 2.

a. The transitional curve between late fetus and post-infant. The curves are considered as the integrals of those drawn in b.

b. Skew logistic growth velocities. There is only one upper asymptote, because separating the sexes in the fetal period was impracticable from the data. The upper asymptote is therefore the fetal common mean growth velocity. The lower asymptote represents the velocities of the post-infantile period. The asymptotes are the derivatives of the straight lines of FIGURES 1 and 2.

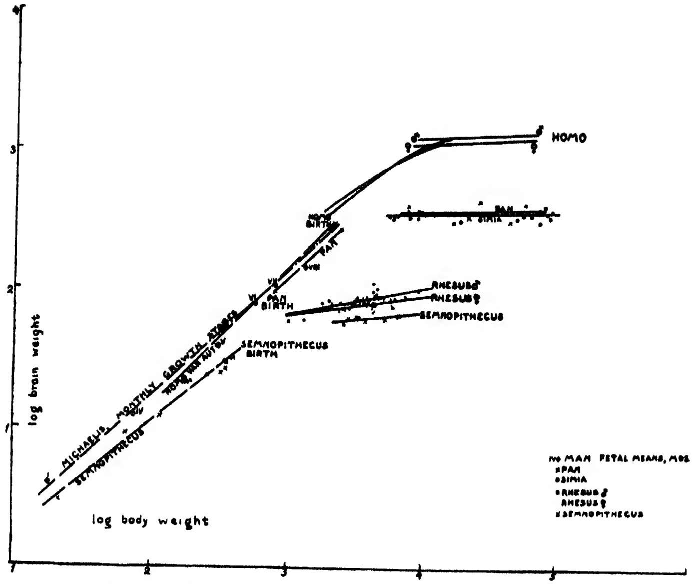


Figure 4. Primate ontogeny. Broken lines: prenatal; solid lines. postnatal.

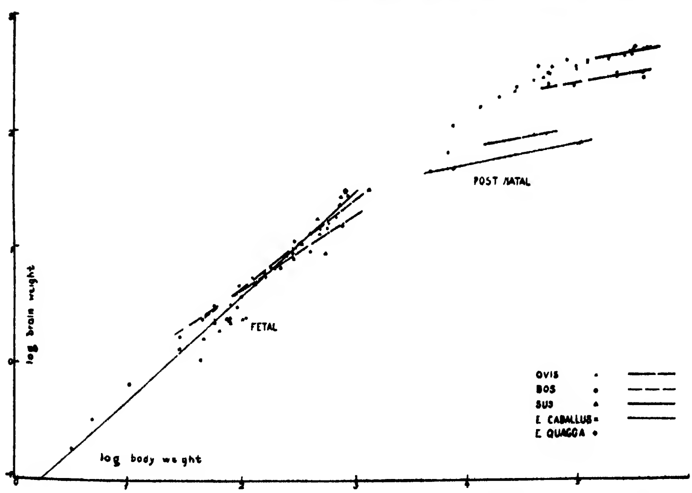


Figure 5. Ungulate ontogeny. Note the parallelisms, but also the different levels and lengths, of the postinfantile lines. This includes the perissodactyls.

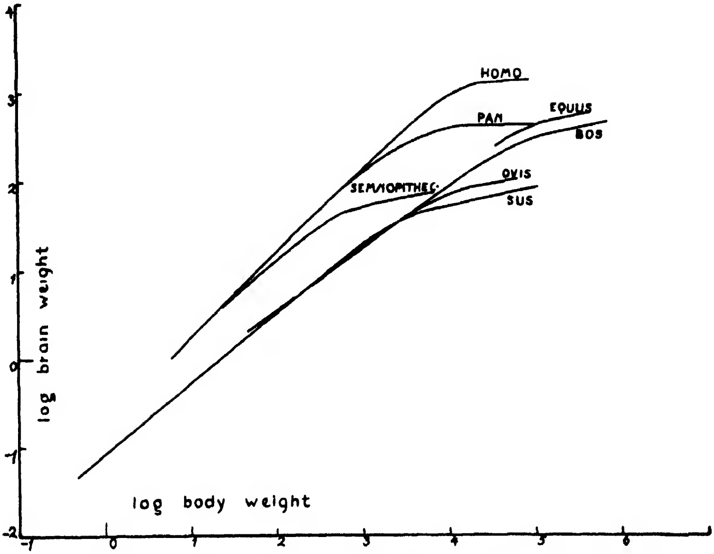


FIGURE 6 A comparison of ontogenetic curves of some primates and ungulates.

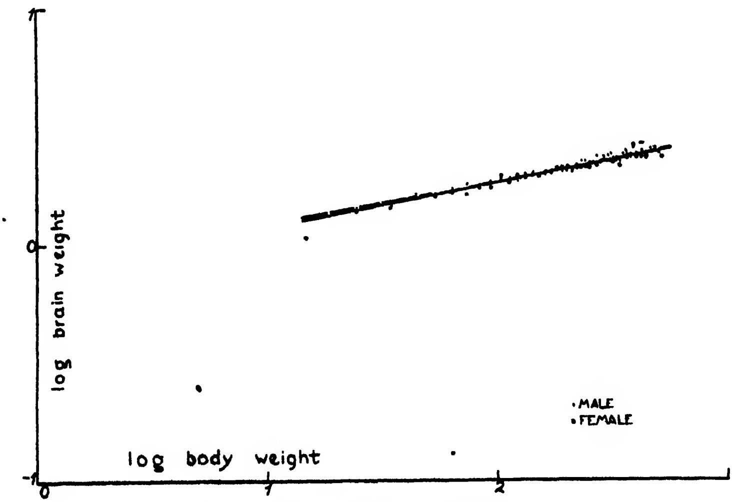


FIGURE 7. Postnatal growth of Norwegian rats.

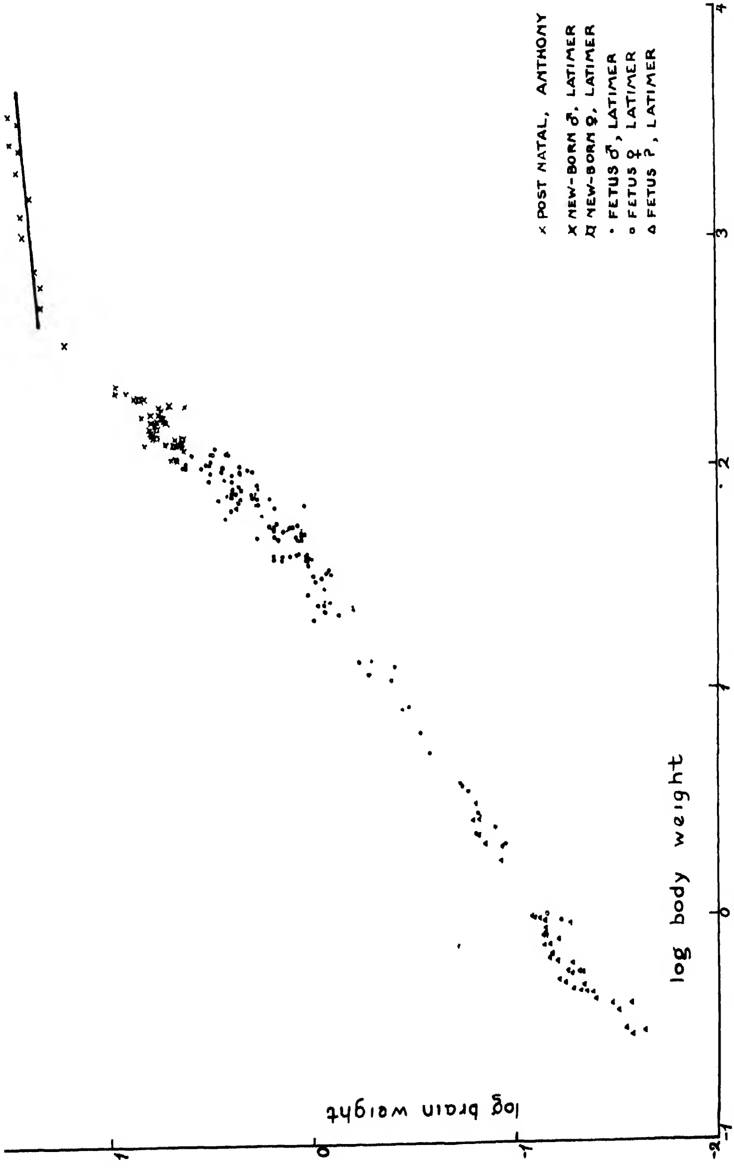


FIGURE 8 Ontogeny of the cat to illustrate that the rectilinear fetal formula probably is only a first approximation The configuration suggests rather a complex logistic with the infantile axis of growth an inclined asymptote

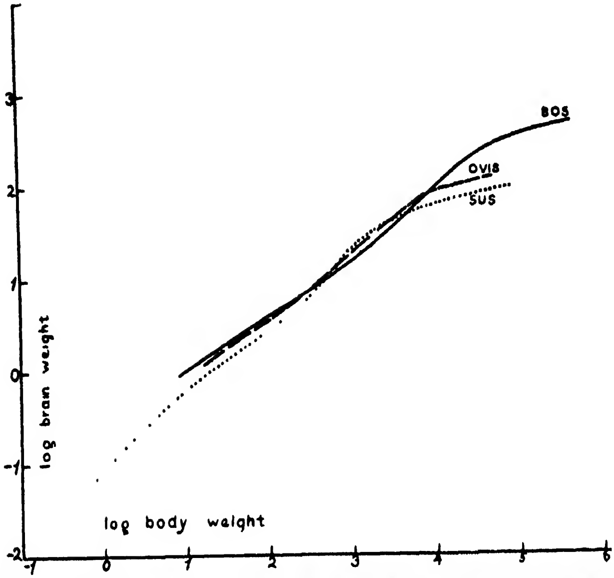


FIGURE 9. The ontogenesis of three artiodactyl genera to illustrate the presumably more correct morphology. This would make of the tabulated fetal formulae but first approximations. Suggested by a comparison of the data with those of **Latimer's** cats (FIGURE 8).

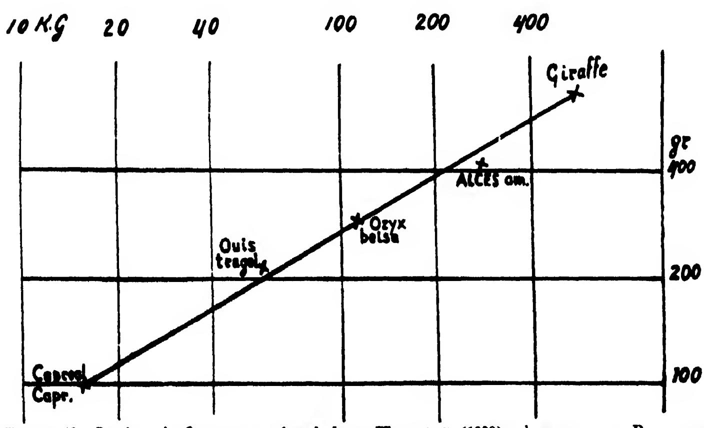


FIGURE 10. Lapicque's figure, reproduced from **Kappers** (1929), showing some Pecora genera fitted to a line $Y = A + 56X$. To illustrate Dubois' hypothesis.

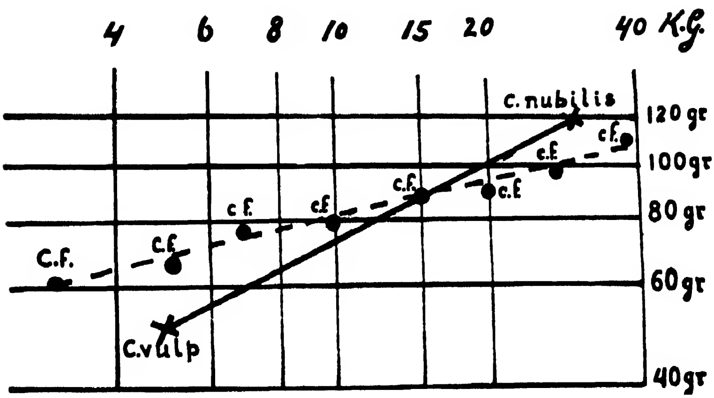


FIGURE 11 Lapicque's figure reproduced from **Kappers**, showing a slope of approximately 25 tted to graded sizes of *Canis familiaris* (c f) and intersecting a slope of 56 which joins two wild)ecies of *Canis*.

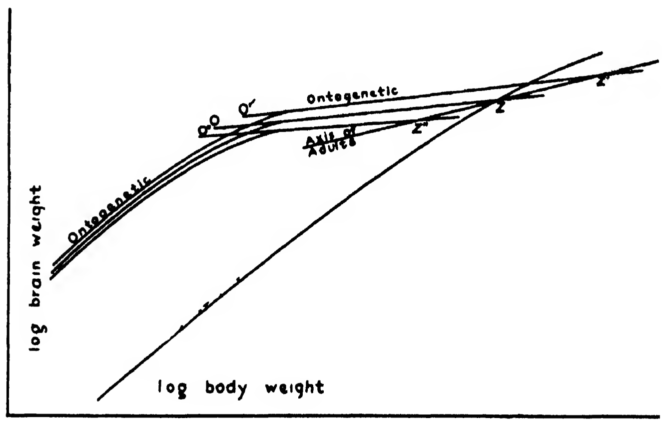


FIGURE 12 A putative explanation of the relationships in a given species or genus between ıtogenetic curves the axis of an adult constellation and a parabolic cephalization exponent hich is comparative anatomic Of the ontogenetic curves, the central one represents the mean of e species It is flanked by individuals respectively bigger and smaller than the average The raight lines, OZ, etc , are postinfantile to adult, the curves being the postnatal infantile The lult population, presumably, is oriented by the line $Z''ZZ'$, it has been drawn to a slope of)out 25 (*cf* **Lapicque**) The comparative-anatomic parabola which is the central theme of ection IV is steeper, and over a considerable stretch may approximate 56 as a slope, or 'rimates) it may be steeper. It is not far from paralleling the fetal ontogenetic curve (see ection V and FIGURE 24).

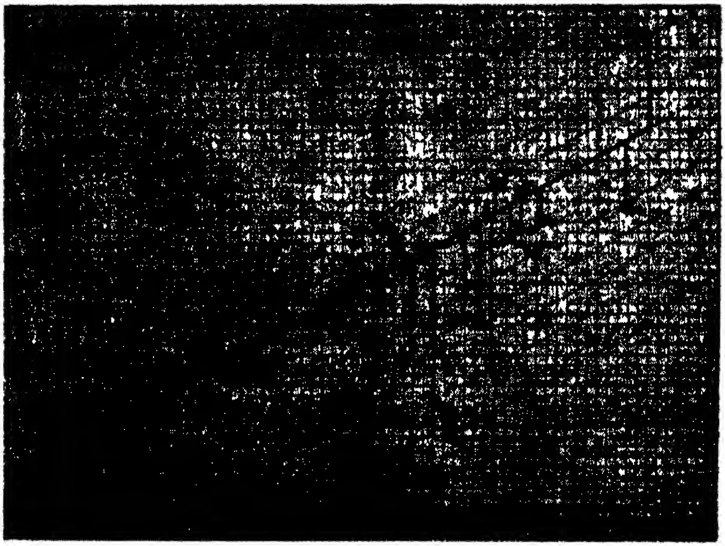


FIGURE 13 The ungulate scheme of **Brummelkamp** (1939c) See TABLE 13

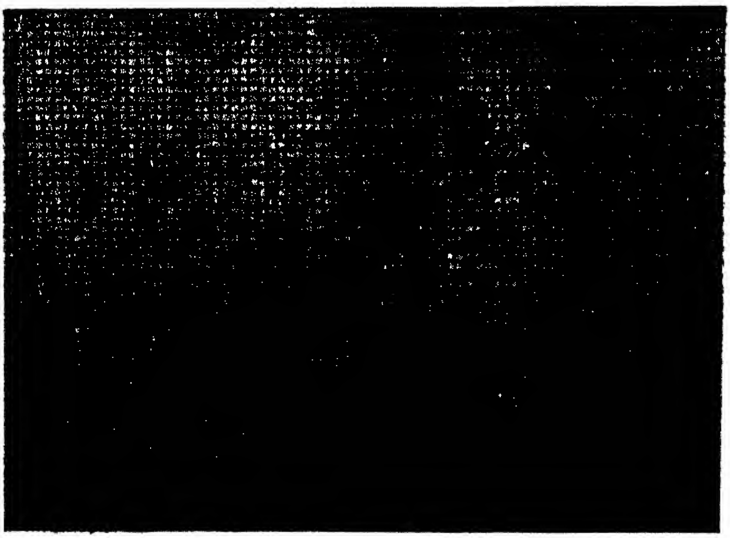


FIGURE 14. The rodent scheme of **Brummelkamp** (1939b) See TABLE 14

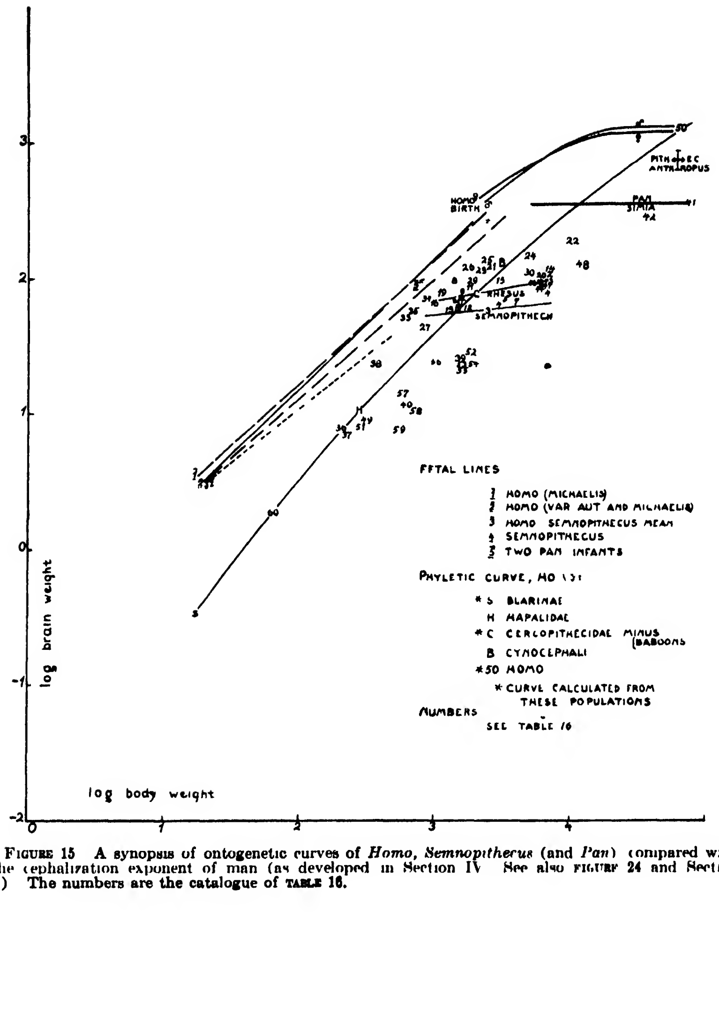


FIGURE 15 A synopsis of ontogenetic curves of *Homo*, *Semnopithecus* (and *Pan*) compared with the cephalization exponent of man (as developed in Section IV See also FIGURE 24 and Section V) The numbers are the catalogue of TABLE 16.

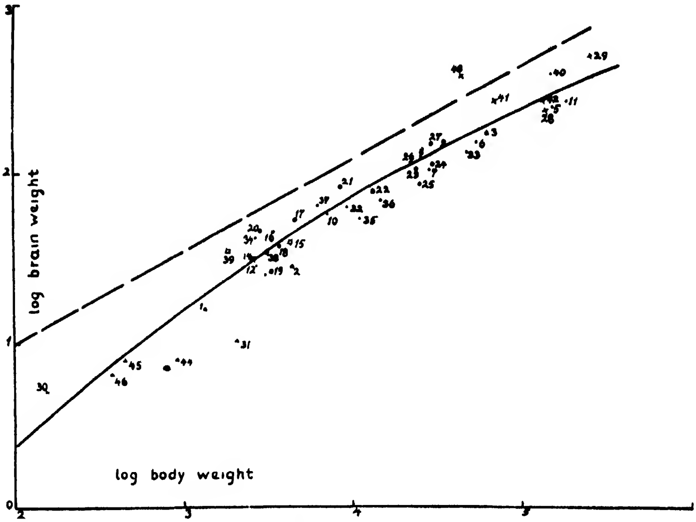


FIGURE 16. Detail of the Carnivora constellation, for comparison with FIGURE 18 (Artiodactyla). The straight, broken line has a slope of .56 and arbitrarily passes through $Y = 0$, $X = 0$. The parabola is formulated in TABLE 26. See TABLE 15 for key to entries.

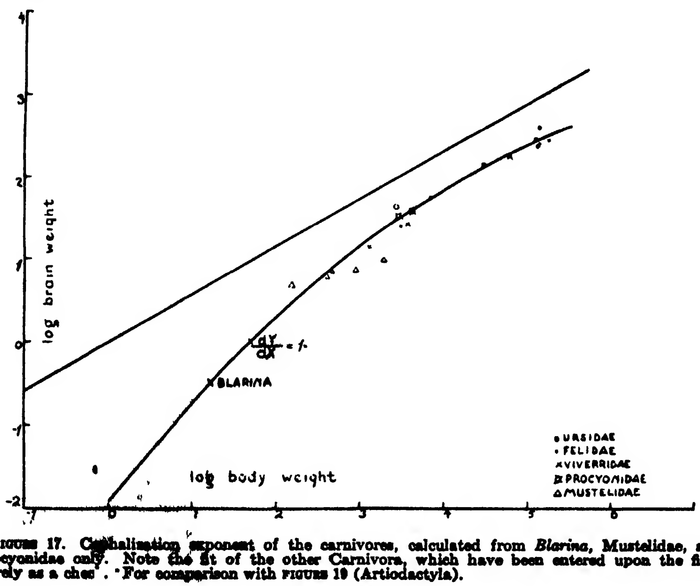


FIGURE 17. Cephalization exponent of the carnivores, calculated from *Blarina*, Mustelidae, and Procyonidae only. Note the fit of the other Carnivora, which have been entered upon the field merely as a check. For comparison with FIGURE 19 (Artiodactyla).

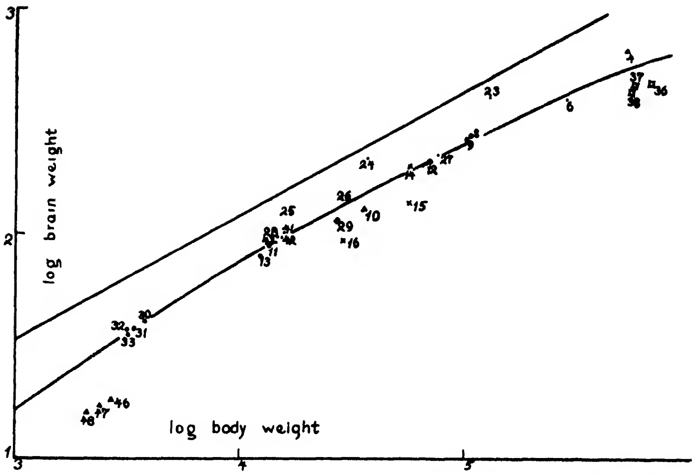


FIGURE 18 Detail from FIGURE 13 The numbering is by **Brummelkamp** (see TABLE 13) The straight line, like the slopes of Brummelkamp, has a slope of .56 and passes, arbitrarily, through $X = 0$, $Y = 0$ The curve has the parabolic formula No. 16 of TABLE 26.

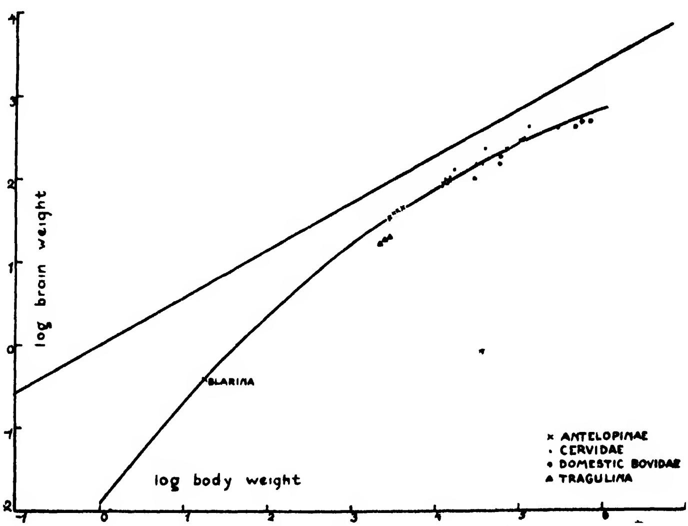


FIGURE 19. The cephalization exponent of the Antelopinae, calculated from *Blarina* and Antelopinae only. Note closeness of fit of the other artiodactyls. For comparison with FIGURE 17 (Carnivora). The rectilinear line slopes at .56, and arbitrarily passes through $Y = 0$, $X = 0$.

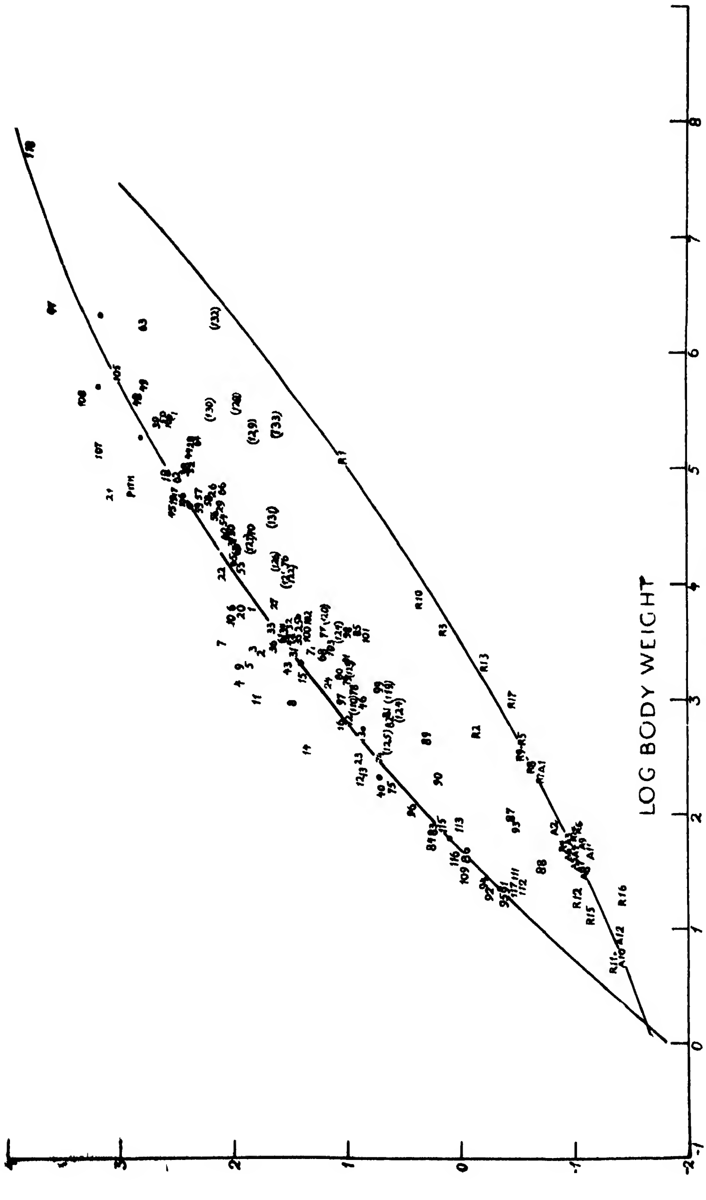


FIGURE 20. See opposite page.

Figure 20 The mean statistical tendency of extant Mammalia (upper curve concave downward) and that of extant Reptilia (lower curve concave upward) Mammalian data from subjoined list reptilian and amphibian from Tables 21 and 22 Plain numbers mammals R numbers reptiles A numbers amphibia The entries in parentheses are extinct mammals, and were not included in the calculation neither were the marsupials although extant and registered in the field The black spots are means (dead centers) of the eleven areas of Table 25

1 *Semnopithecus*
2 *Cercopithecus*
3 *Chlorocebus*
4 *Cercocebus*
5 *Macacus*
6 *Cynopithecus*
7 *Cynocephalus*
8 *Mycetes*
9 *Ateles*
10 *Lagothrix*
11 *Cebus*
12 *Hapale*
13 *Iacchus*
14 *Midas*
15 *Lemur*
16 *Nycticebus*
17 *Simia*
18 *Gorilla*
19 *Pan*
20 *Hylobates*
21 *Homo*
22 *Symphalangus*
23 *Oedipomidas*
24 *Genetta*
25 *Ichneumia*
26 *Crocuta*
27 *Felis*
28 *Felis*
29 *Felis*
30 *Felis*
31 *Potos*
32 *Procyon*
33 *Vulpes*
34 *Urocyon*
35 *Otocyon*
36 *Thos*
37 *Canis*
38 *Ursus*
39 *Thalarctos*
40 *Mustela*
41 *Mephitis*
42 *Nasua*
43 *Potos*
44 *Melursus*
45 *Helarctor*
46 *Putorius*
47 *Elephas*
48 *Camelus*
49 *Giraffa*
50 *Equus*
51 *Alces*
52 *Oryx*
53 *Antilope*
54 *Antil capra*
55 *Antilope*
56 *Gazella*
57 *Ovis*
58 *Sus*
59 *Cervus*
60 *Rupicapra*
61 *Cephalophus*
62 *Damaliscus*
63 *Hippopotamus*
64 *Tapirus*
65 *Capreolus*
66 *Phacochaerus*
67 *Tayassu*
68 *Tragulus*
69 *(Zebra)*
70 *Hydrochaerus*
71 *Dasyprocta*
72 *Sciurus*
73 *Sciurus*
74 *Sciurus*
75 *Sciurus*
76 *Hystrix*
77 *Lepus*
78 *Lepus*
79 *Oryctolagus*
80 *Pteromys*
81 *Cavia*
82 *Cavia*
83 *Dipus*
84 *Sciuropterus*
85 *Lagostomus*
86 *Arvicola*
87 *Arvicola*
88 *Mus*
89 *Mus*
90 *Mus*
91 *Mus*
92 *Mus*
93 *Mus*
94 *Cricetulus*
95 *Mus*
96 *Geomys*
97 *Perod pus*
98 *Capromys*
99 *Fiber*
100 *Tamanduas*
101 *Dasypus*
102 *Choloepus*
103 *Bradypus*
104 *Erignathus*
105 *Odobenus*
106 *Phoca*
107 *Phocaena*
108 *Delphinapterus*
109 *Desmodus*
110 *Pteropus*
111 *Rhinolophus*
112 *Vespertilio*
113 *Tupaja*
114 *Erinaceus*
115 *Talpa*
116 *Scalopus*
117 *Blarina*
118 *Balaenaptera*
119 *Didelphis*
120 *Onychogale*
121 *Macropus*
122 *Petrogale*
123 *Trichosurus*
124 *Trichosurus*
125 *Dasyurus*
126 *Thylacinus*
127 *Mesohippus*
128 *Paleosyops*
129 *Anoplotherium*
130 *Moeritherium*
131 *Diplobune*
132 *Uintatherium*
133 *Coryphodon*

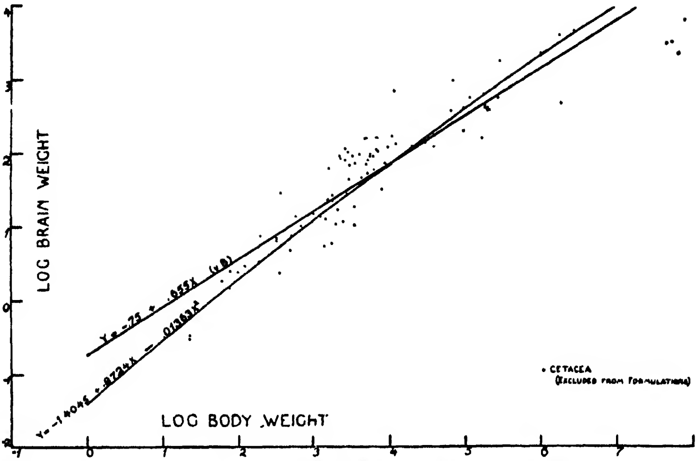


FIGURE 21. Mammalian constellation of von Bonin, fitted with his regression line, and a parabola, calculated as *a*, explained in the text. Compare with FIGURE 20. The straighter character of the parabola in the present case is due, in part at least, to the exclusion of the giant Cetacea from the data This was not done in FIGURE 20.

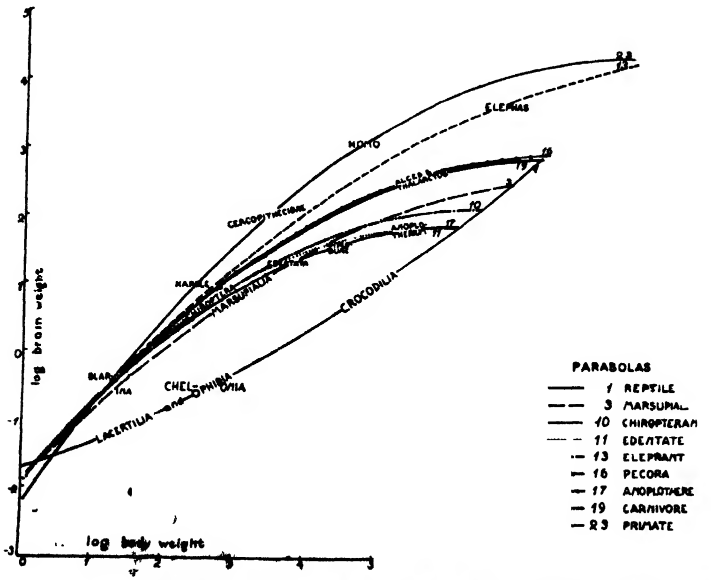


FIGURE 22. Some exponents of cephalization drawn from formulae of TABLE 26, and numbered accordingly. The parabolas pass indefinitely; a very few genera, orders, etc., have been placed in the region where they occ [illegible] visualization.

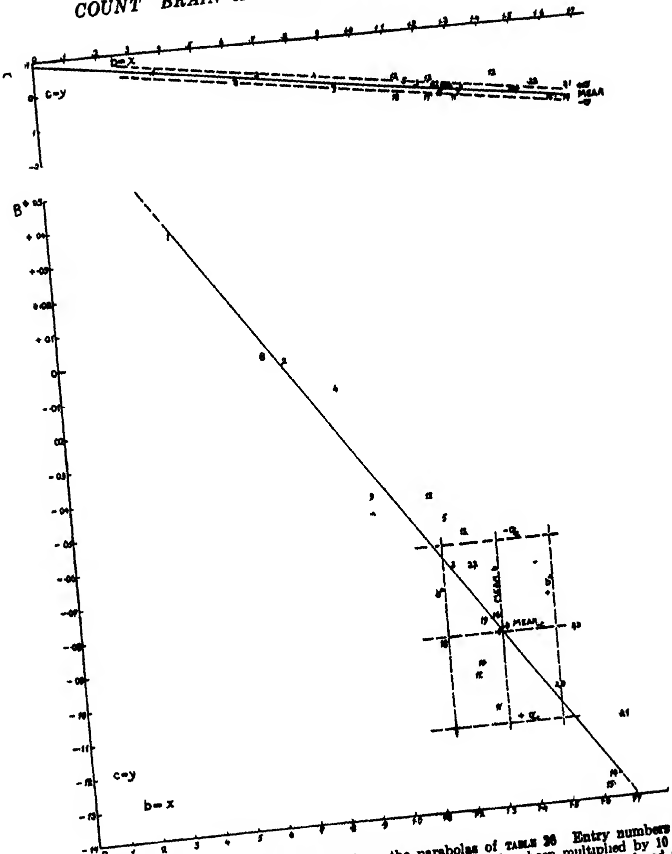


FIGURE 23 Plot of b as x and of c as y, from the parabolas of TABLE 26 Entry numbers accordingly In A the scales are the same in B the ordinate scale has been multiplied by 10 In A, the broken lines are the $\pm\sigma$ limits if all points are projected on to a common line placed normal to the line of the mean In B, the broken-line rectangle limits the $\pm\sigma$ of mean b and of mean c, respectively This illustrates the degree of "typicalness" of each of the parabolas of TABLE 26, when the mean of all *surviving* cephalisation exponents, regardless of the absolute brain and body weights of the genera that they characterise is taken as standard

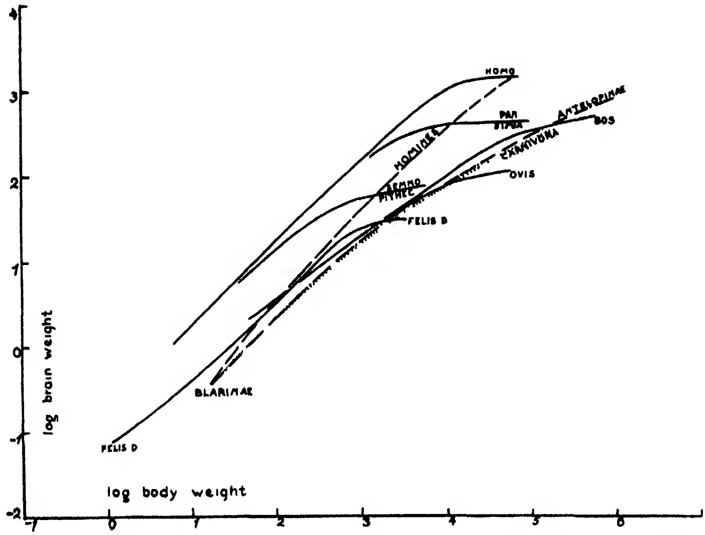


FIGURE 24 Synoptic comparison of ontogenetic and comparative-anatomic curves of some genera-orders Solid lines ontogenetic, broken lines comparative-anatomic Note parallelisms between fetal and comparative-anatomic trends, yet with fetal always showing a brain precocity Compare with FIGURE 12.